DOCUMENT RESUME

ED 387 524 TM 023 908

AUTHOR	Frechtling, Joy A., Ed.
TITLE	Footprints: Strategies for Non-Traditional Program Evaluation.
INSTITUTION	Westat, Inc., Rockville, MD.
SPONS AGENCY	National Science Foundation, Arlington, VA. Directorate for Education and Human Resources.
REPORT NO	NSF-95-41(new)
PUB DATE	Jan 95
CONTRACT	SED-925569
NOTE	164p.; Papers prepared for a conference on nontraditional evaluation methodologies convened by the National Science Foundation in July 1993.
PUB TYPE	Collected Works - General (020)
EDRS PRICE	MF01/PC07 Plus Postage.
DESCRIPTORS	*Evaluation Methods; Evaluation Utilization; *Information Dissemination; *Innovation; Mathematics Education; *Program Evaluation; *Research and Development; Science Education
IDENTIFIERS	*Impact Evaluation; *National Science Foundation; Stakeholder Evaluation

ABSTRACT

Papers in this collection explore alternative and nontraditional approaches to evaluation. They provide options, speculations, and propositions that affect each thinker's ideas on how to trace the impact of National Science Foundation-supported programs. Organized around a central theme of footprints as evidence of a program's impact, the papers include: (1) "The Use of Science and Mathematics Education Indicators and Studies: A Briefing" (Robert F. Boruch and Erling Boe); (2) "Searching Near, Far, and Wide: A Plan for Evaluation" (Sylvia T. Johnson); (3) "New Methods for Evaluating Programs in NSF's Division of Research, Evaluation, and Dissemination" (Robert K. Yin with commentary by Valena White Plisko, David Jenness, and Malcom Phelps); (4) "Considerations for the Evaluation of the National Science Foundation Programs" (Richard T. Hezel); (5) "Communicating the Value of the National Science Foundation's Contributions to Research and Innovative Technical Applications for Mathematics and Science Education" (Norman L. Webb); (6) "Footprints on Surfaces: A Nontraditional Approach to Evaluation of National Science Foundation Programs" (M. Christine Dwyer with commentary by Robert Mac West and Senta Raizen); (7) "Conceptual Underpinnings for Program Evaluation of Major Public Importance: Collaborative Stakeholder Involvement" (Zoe A. Barley and Mark Jenness); (8) "The Virtual Reality of Systemic Effects of NSF Programming on Education: Its Profession, Practice, Research, and Institutions" (Robert E. Stake with commentary by Eleanor Chelimsky and David B. Rymph); (9) "Overview" (Michael Scriven); and (10) "Footprints: A Search for New Strategies for Evaluating EHR [Education and Human Resources] Programs" (Laure Sharp and Joy Frechtling). References follow each paper. (SLD)

TM

ED 387 524

DIVISION OF RESEARCH, EVALUATION AND DISSEMINATION

FOOTPRINTS:

Strategies for Non-Traditional Program Evaluation

Sponsored by the National Science Foundation

Coordinated by Westat, Inc.

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it
☐ Minor changes have been made to improve reproduction quality
- Points of view or opinions stated in this document do not necessarily represent official OERI position or policy

A RED - sponsored Monograph on Evaluation

National Science Foundation
Directorate for Education and Human and Resources

TM023908

BEST COPY AVAILABLE

The Foundation provides awards for research in the sciences and engineering. The awardee is wholly responsible for the conduct of such research and preparation of the results for publication. The Foundation, therefore, does not assume responsibility for the research findings or their interpretation.

The Foundation welcomes proposals from all qualified scientists and engineers, and strongly encourages women, minorities, and persons with disabilities to compete fully in any of the research and related programs described here.

In accordance with federal statutes, regulations, and NSF policies, no person on grounds of race, color, age, sex, national origin, or disability shall be excluded from participation in, denied the benefits of, or be subject to discrimination under any program or activity receiving financial assistance from the National Science Foundation.

Facilitation Awards for Scientists and Engineers with Disabilities (FASED) provide funding for special assistance or equipment to enable persons with disabilities (investigators and other staff, including student research assistants) to work on an NSF project. See the program announcement or contact the program coordinator at (703) 306-1636.

Privacy Act and Public Burden

Information requested on NSF application materials is solicited under the authority of the National Science Foundation Act of 1950, as amended. It will be used in connection with the selection of qualified proposals and may be used and disclosed to qualified reviewers and staff assistants as part of the review process and to other government agencies. See Systems of Records, NSF-50, "Principal Investigator/Proposal File and Associated Records," and NSF-51, "Reviewer/Proposals File and Associated Records," 56 Federal Register 54907 (Oct. 23, 1991). Submission of the information is voluntary. Failure to provide full and complete information, however, may reduce the possibility of your receiving an award.

The public reporting burden for this collection of information is estimated to average 120 hours per response, including the time for reviewing instructions. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Herman G. Fleming, Reports Clearance Officer, Division of CPO, NSF, Arlington, VA 22230; and the Office of Management and Budget, Paperwork Reduction Project (3145-0058), Wash., D.C. 20503.

The National Science Foundation has TTD (Telephonic Device for the Deaf) capability, which enables individuals with hearing impairment to communicate with the Foundation about NSF programs, employment, or general information. This number is (703) 306-0090.

FOOTPRINTS: Strategies for Non-Traditional Program Evaluation

Edited by:

Joy A. Frechtling
Westat, Inc.

Prepared for:

Susan Gross, Program Officer
Division of Research, Evaluation and Dissemination
Directorate for Education and Human Resources
National Science Foundation
Arlington, VA

January, 1995

Prepared by Westat, Inc., Rockville, MD,
for the Division of Research, Evaluation and Dissemination (RED),
Education and Human Resources (EHR) Directorate,
Susan Gross, Program Officer

The conduct of this study and preparation of this report were sponsored by the National Science Foundation, Directorate for Education and Human Resources, Division of Research, Evaluation and Dissemination, under Contract No. SED-925569. The ideas in the papers and discussions presented at the conference are those of the participants alone and do not necessarily reflect the policies or opinions of the institutions or agencies with which they are affiliated.

Table of Contents

Introduction

Daryl E. Chubin
Foreword v

Joy A. Frechtling
The Search for Footprints: Nontraditional Approaches to Evaluating NSF's Programs 1

Susan Gross
Dear Reader 3

The Papers and Discussants

Robert F. Boruch and Erling Boe
The Use of Science and Mathematics Education Indicators and Studies: A Briefing 7

Sylvia T. Johnson
Searching Near, Far, and Wide: A Plan for Evaluation 15

Robert K. Yin
New Methods for Evaluating Programs in NSF's Division of Research, Evaluation, and Dissemination 25

Valena White Plisko 37

David Jenness 39

Malcom Phelps 43

Richard T. Hezel
Considerations for the Evaluation of the National Science Foundation Programs 45

Norman L. Webb
Communicating the Value of the National Science Foundation's Contributions to Research and Innovative Technical Applications for Mathematics and Science Education 53

M. Christine Dwyer
Footprints on Surfaces: A Nontraditional Approach to Evaluation of National Science Foundation Programs 75

Robert Mac West 91

Senta Raizen 93

Zoe A. Barley and Mark Jenness
Conceptual Underpinnings for Program Evaluation of Major Public Importance: Collaborative Stakeholder Involvement 97

Robert E. Stake
The Virtual Reality of Systemic Effects of NSF Programming on Education: Its Profession, Practice, Research, and Institutions 107

Eleanor Chelimsky 127

David B. Rymph 129

Table of Contents (continued)

Concluding Comments

Michael Scriven
Overview *131*

Laure Sharp and Joy Frechtling
Footprints: A Search For New Strategies For Evaluating EHR Programs *139*

Index *155*

Foreword

Daryl E. Chubin
Director, Division of Research,
Evaluation, and Dissemination

A major responsibility of the Division of Research, Evaluation and Dissemination (RED) is to provide conceptual and technical assistance for the evaluation of projects and programs throughout NSF's Directorate for Education and Human Resources (EHR). The "Footprints" conference was organized in the spirit of "research on practice". We called on innovative thinkers and seasoned practitioners in the educational research community to propose fresh ideas and new methodologies that might inform the design of EHR evaluations. The result is this "Footprints" publication.

As a conference participant (in my waning pre-NSF days) and reader of the papers and discussions reported in this volume, I was especially struck by the call for two tasks which we in RED have begun to undertake:

- Identify and differentiate the audience for EHR program and project evaluations. Our immediate audience for a given evaluation is likely to include program managers and division directors, but we must also consider the information needs of the broader federal community, given its emerging emphasis on evaluation.
- Develop a clear policy with respect to the link between evaluation and dissemination. We see dissemination as a simple concept that denotes a range of activities as one of our primary responsibilities to EHR. We are committed to sharing widely research findings that can be translated into innovative classroom practice and help us achieve national goals for the improvement of mathematics and science education, for all students.

We hope that by building on this volume, we can expand and fine-tune our repertoire of evaluation strategies, and determine better ways of matching different evaluation needs with different approaches. My EHR colleagues and I see this as a major way of contributing to the success of reform initiatives nationwide. We cannot do this alone. Therefore, I welcome your comments on this volume and RED's other evaluation products.

Finally, I am grateful to Westat's Laure Sharp and Joy Frechtling, and to Susan Gross of the RED Evaluation staff for bringing the "footprint" metaphor to practical function. I am privileged to be positioned within NSF so as to apply the lessons of this conference to EHR's formidable schedule of program evaluations.

The Search For Footprints: Nontraditional Approaches To Evaluating NSF's Programs

Joy A. Frechtling

The National Science Foundation (NSF) supports a number of programs that are designed to produce state-of-the-art research and innovative technical applications for mathematics and science education. Projects funded under these programs vary widely in their scope, size, and duration. Some are one-of-a-kind efforts, designed to investigate a new approach, theory, or technology. Some may be part of a stream of research, involving projects that build on each other to create a comprehensive model or those that move from theory to practice. Still others represent cooperative ventures that blend the resources of NSF with those of other funding agencies to address issues of joint interest.

While the peer review process for selection of grantees provides one important type of evaluation of NSF's programs (in the sense of quality control over what is supported), NSF, like other government and private agencies, also needs to conduct more formal program evaluations—evaluations that can be used to document the impacts and, as relevant, the shortcomings of its programs. Quality control needs to be supplemented by quality review.

However, evaluating programs such as the ones described above is neither easy nor straightforward. Traditional educational evaluation strategies that have been useful in evaluating programs that support the delivery of new services, instructional strategies, or curricula (the most familiar and widespread evaluation challenge) are not directly applicable to the majority of the research-oriented, groundbreaking inquiries that make up the portfolios of many of the Foundation's efforts. Further, the kinds of program impacts that can and should be expected of many NSF programs differ in some important ways from those typically considered where service delivery projects are the focus of study. For example,

- Traditional educational evaluations seek to attribute any impacts found to a single source, be it a support program such as Chapter 1 or a classroom intervention such as cooperative learning. For many of the programs at NSF, drawing such uni-dimensional causal statements is unlikely or impossible.
- Traditional educational evaluations have relied almost entirely on quantitative data or on counts of events. For many of the programs funded by NSF simple counts are misleading; a single successful project may justify the entire research investment, and use of quantitative indicators may exclude important areas for which no appropriate quantitative measures exist.
- Traditional educational evaluations of programs in the education sector have given priority to measures of student achievement as the impact measure of greatest concern. For many NSF programs, student achievement is an inappropriate measure either because of the nature of the research itself or the fact that any impacts on students would not be expected in the short run.

Recognizing this lack of alignment between traditional evaluation models and the nature of the programs that NSF needs to examine, the Division of Research, Evaluation, and Dissemination commissioned a series of papers designed to explore alternative, nontraditional approaches to evaluation. The goal is to "stretch our minds" with regard to evaluation and to explore new options, rather than to stipulate new prescriptions. This NSF project, dubbed "Footprints," is an attempt to examine the impacts of funding programs that have been part of NSF's repertoire for a number of years and to assess the impressions they have made on the field, on scholarship, on other institutions, and on practice. This monograph presents the results of that project.

In reading the papers, it is important to keep in mind that they are not evaluations of any particular program or programs. Nor are they, in many cases, fully developed designs that could be adopted and used tomorrow or next week. Rather, they are options, speculations, and propositions that represent each thinker's ideas on how one might trace the impact of NSF's programs of support. Further, while designed with NSF programs specifically in mind, the approaches should provide food for thought for other institutions and agencies faced with similar evaluation challenges.

The papers have been solicited from a diverse group of thinkers who approach the evaluation task from both differing backgrounds and philosophies. And, because they were encouraged to think broadly in constructing their interpretations, they have produced conceptualizations of "nontraditional" that vary along a number of different dimensions.

- Some authors have emphasized the need for nontraditional evidence, indicators of program success that vary significantly from the student achievement indicators that have characterized more traditional studies. In line with this, several authors have looked for footprints in terms of effects on actual practice, on accepted models of learning, on methodologies, and even on policy.
- Some authors have stressed the development of nontraditional methodologies, supplementing the quantitative approach with one that relies more, or even exclusively, on qualitative inquiry. In fact, almost all the papers include qualitative analysis to some extent or another.
- Another dimension of difference is that of the role of the stakeholder, as opposed to the professional investigator, as the generator of hypotheses and the discoverer of impacts. Following recent trends in evaluation, almost all the papers underline the importance of stakeholder involvement—especially in understanding program goals and objectives. Some go even further and see the role of stakeholders as central to the whole evaluation enterprise.
- The papers also differ in the extent to which program evaluation is seen as an aggregation of the evaluations of different projects versus an evaluation of the program as a whole. Those who fall into the first camp seem to feel that the same outcomes that are used for assessing project success can, and should be somehow aggregated to assess program success. Others seems to follow the old saying that "the whole is greater than the sum of its parts" and seek other sources of evidence.
- Finally, the papers also differ in what could be called the "level of maturity" of the proposals being offered. Some could probably be implemented tomorrow, or at least next month, if NSF chose to do so. Others are more preliminary and will need considerably more thought and development before it is possible to assess their efficiency. These provide a core of ideas for new research on evaluation methodologies should NSF or some other agency choose to move in that direction.

Also included are a series of "reaction statements." These are not fully developed papers as such but, rather, brief statements offered in response to some of the ideas expressed. Some provide challenges to the authors; others are endorsements of an idea or point of view. The final responses attempt to put the ideas into perspective and provide suggestions for the next steps that NSF might take.

Susan Gross
National Science Foundation

Dear Reader:

The papers and discussions contained in this monograph were prepared for a conference on non-traditional evaluation methodologies that was convened by the National Science Foundation in July 1993.

NSF embarked on this project because of a need to evaluate several of its programs that were not structured in the typical service delivery model. The programs support research projects and special studies that are designed to shed light on what we know about the teaching and learning of science and mathematics. Four NSF programs were the focus of the commissioned papers:

Research in Teaching and Learning
RTL supports projects which investigate how individuals and groups learn, teach, and work effectively in complex, changing environments.

Applications of Advanced Technologies
AAT supports research, development, and proof-of-concept projects that address issues at the forefront of technology applications to learning and teaching in science and mathematics.

Studies
The Studies Program supports research projects on significant factors, trends, and practices in education, with an emphasis on their policy application.

Indicators
The Indicators Program supports studies that provide statistical information about the status of mathematics and science learning.

In my introductory remarks at the conference, I attempted to illustrate what we are looking for by use of the "footprint" metaphor. The metaphor arose from preliminary discussions concerning the four NSF programs in need of evaluation. Evaluation of these programs presented a challenge; we needed to find evidence that the programs were leaving "footprints in the sand" of mathematics and science education in the nation. Thus, the conference became known within NSF and among the authors as the "Footprints" Conference. The following ramblings are the remarks I made at the conference. The illustrations shown here were actually light-hearted computer art that was prepared for the conference—alas, they lose a bit in the translation. They are included here at the suggestion of several conference participants who felt they helped establish a focus or context for the day. I hope they work as well in print.

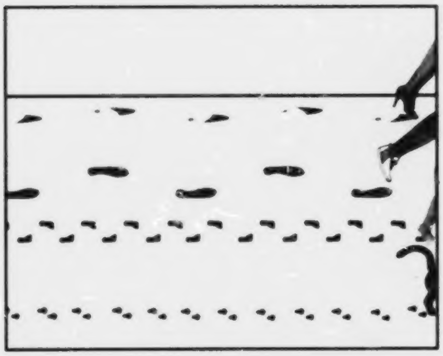
Susan Gross
Program Officer
National Science Foundation
Division of Research, Evaluation, and
Dissemination

Remarks

The central theme in the papers that we commissioned is Footprints.

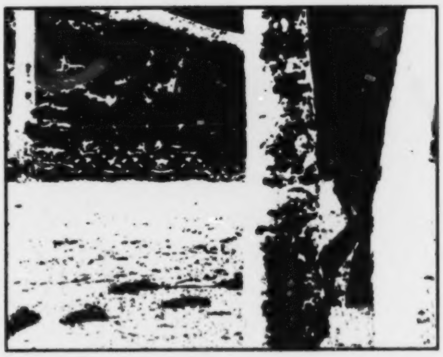
Footprints can be viewed as the evidence of a program's impact.

Examples include: evidence that the program has had an effect on mathematics or science education; evidence that the results obtained from one or more projects funded by a program are disseminated and used elsewhere.



Footprints come in various shapes and sizes.

We should look for many types of programmatic effect, for example, changes in how we think about teaching and learning; evidence that the latest research is considered when teacher training programs are planned; examples of how the latest developments in technology are used in classroom instruction. Different types of evidence are appropriate for different types of programs. We would hope to see payoffs of research programs affecting teacher training, classroom instruction, and student learning. The production of statistical data reports, on the other hand, might result in changes in national or state policy.



Some Footprints will last a long time. This can be both good and bad.

When something worthwhile has been accomplished, it should be disseminated, replicated, and thoroughly examined and understood. However, there is the danger of a good thing hanging around too long and becoming out of date or no longer the best thinking. LOGO is an example of a computer language that served its purpose and is no longer considered state-of-the-art. Emphasis on basic skills instruction to the exclusion of higher order thinking and solving of complex problems is no longer considered the best educational approach.

The surface in which Footprints are left is important.

If the surface is not prepared adequately, the findings will be washed away. A properly prepared surface will allow lasting impressions to be made. This means that stakeholders (e.g., program planners, decision-makers, project PIs) should be involved in planning the evaluation so they will be accepting of the results.

We need to know where to look for Footprints.

How do we know we have collected all the evidence? Where are the likely places to look for missing evidence? For example: What are the untouched areas of research? What is not being done or is being done ineffectively? Are there key target groups that are not being served or are being served inadequately? What rival hypotheses can we formulate, e.g., where would we have been if this program did not exist?

We need to know when a Footprint has outlived its usefulness.

Yesterday's goal for education reform may no longer be a goal because we have moved beyond it. We need to be vigilant in retiring or making extinct old goals and adopting new ones that move us to higher standards of excellence. We must examine with regularity statistical indicators that we use to assess the health of the nation in mathematics and science education. An indicator can lose meaning because the nation has attained it, or because people work toward it as the end product rather than as the means to a larger end.

BEST COPY AVAILABLE

The Use Of Science And Mathematics Education Indicators And Studies: A Briefing

Robert F. Boruch and Erling Boe
University of Pennsylvania

This briefing concerns the "footprints" that might be made by an array of projects sponsored by the National Science Foundation (NSF). "Footprints" here means (roughly) traces of whether and how the results of the projects were used. The object is to speculate on what uses of data or studies are worth looking for and why, and how one might discern them.

The target research of interest includes the statistical surveys sponsored by the NSF's Indicators Program, such as the Third International Study of Mathematics and Science. It includes policy-related work supported by the Studies Program, such as the examination of test and textbook contents and how these relate to the higher order thinking skills of students done by Madaus and colleagues.

This paper summarizes a longer report on the topic and capitalizes heavily on information supplied by NSF. Foundations such as the Rockefeller Foundation and agencies, the National Center for Education Statistics (NCES), and the Planning and Evaluation Service of the U.S. Department of Education have also posed questions about the value of the studies they sponsor. Their experience is also exploited here.

A major premise underlying this effort is that the data and studies on science and mathematics education produced under NSF sponsorship should be "useful." This premise is critical in that some research products are important in the long run by a variety of standards but can be regarded as useless by a variety of other standards. The premise is fundamental, but its import is debatable at the margin.

"A major premise underlying this effort is that the data and studies on science and mathematics education produced under NSF sponsorship should be 'useful.'"

Conclusions are framed in terms of the lessons learned from contemporary social research on the use of data and policy studies. These conclusions cover essential formalities such as definition of the "use" of a data set or study, common methods of tracking use, the uncommon and underexploited methods of tracking, and planning for enhanced data or study use.

1. It is essential to define what is meant by the "use" of information and to distinguish among types of use. It is essential also to define the initial conditions and context of use.

Statistical indicators and studies of science and mathematics education may be "used" in the senses of (a) being recognized or seriously considered, (b) informing decisions, and (c) leading to actions. Making plain what is meant by data or study use is essential for program monitoring, of course, and can help to prevent egregious argument about what has been useful.

Different kinds of use must usually be discerned in different ways. The NSF-sponsored data on the U.S. rank in science or mathematics education (SME) relative to other countries have arguably influenced public debate regardless of any specific corporate or public decisions. The debate is traceable by examining public and professional press coverage. This debate arguably informed decisions to focus U.S. education goals on SME and were arguably followed by action—appropriation of funds for SME. The extent to which the debate informed decisions may be traceable through legislative hearings. The extent to which these decisions led to action may be discernible by observing changes in level of Federal appropriations for science and mathematics education research.

"... conventional citation counts fail to recognize influential studies that are not reported in journals."

Definition also means specifying initial conditions, context, and constraints. In the case of NSF and other foundations that sponsor the production of data, the initial conditions include institutional memory that is limited by staff rotation, formal data banks that are limited by resources, a basic science culture that puts priority on "interesting and important" rather than on "useful and important," and a foundation stress on "push the cargo out and fly on." At the individual level, the initial conditions include the roles of program staff members and their relations with aspiring principal investigators, the limits on the role of each, and the subcultures in which each operates.

2. It is easy to identify methods of tracking the uses to which statistical data and studies are put, but the methods are not commonly exploited by foundations that sponsor research.

A variety of ways have been invented to register the production and use of a data set or study. The common ones include the following:

- Counts of the publication of study results, especially publications in refereed journals and high quality books, coupled with estimates of how many scholars on average read how many articles in the relevant journals;
- Awards to a person or group, especially those made by independent professional organizations, for scholarly products generated through the study or data set;
- Popular press or media coverage of the study or its product, e.g., op-ed articles;
- Presentations in professional forums and especially in public forms in which decisions about exploiting the data are made; and
- Citation counts, notably of journal articles, books, or presentations that depend on the data set or study of interest.

Each has merit. Science journal citation counts, for instance, are an inexpensive device for learning whether certain academic audiences attend to the study. Each device, too, has shortcomings. Citation counts that focus on scholarly journals are arguably ineffective for important potential users, such as policy makers. In any event, conventional citation counts fail to recognize influential studies that are not reported in journals.

These methods of tracking the production of data and studies and their use have been identified elsewhere and are, indeed, employed to gauge an entity's performance. For instance, the U.S. General Accounting Office's (GAO) Annual Report to the Congress has in recent years included the number of studies undertaken, the number of GAO reports produced, and the incidence of congressional testimony by GAO staff.

At GAO, output indicators such as production of reports are almost inseparable from "use" indicators because most of such reports are requested by Congress and presumably used by the requestor. Nonetheless, where evidence is sufficient, the Annual Report also provides narrative information on the consequences of particular studies, e.g., reduction in fraud, waste, or abuse. Any Federal agency such as NSF, that produces studies and data that are supposed to be useful, might produce a similar report.

These simple methods are uncommon in that they are not systematically exploited by foundations or other government agencies that produce studies. NSF, for example, has no archive of publications produced by the researchers that it sponsors; it is not clear that NSF has the resources for an archive. In any event, a custom would need to be invented to assure that researchers send publications and presentations to NSF to build such an archive; a mechanism would have to be created to assure that the archive is used.

3. Statistical data and study results are woven into applied research and analysis, often in nonobvious ways. It is important to take into account imperfect recognition of a data set or study and to understand data filters and intermediary users of the information.

Low-level, persistent use of information can be important. But traces of it are often weak. Popular press reports, for example, often do not identify properly a study's sponsor, the research entity, or study's name. Refereed scholarly journals only at times properly acknowledge the specific data sets that were used in a publication.

More generally, data and studies pass through a variety of filters or, as Chris Dwyer calls them, intermediaries. The Congressional Budget Office (CBO) may reanalyze data set X, analyzed by professor Y, both sponsored by NSF. The GAO might cite the CBO's work without a reference to the NSF sponsor or the original analyst. This implies that the "reader" who is hired to track the uses of a study, or the electronic scanning strategies that are invented to track data use, must be flexible in going beyond a given user of information to the preceding one.

Identifying instances in which a data set or study is used, in literature that ranges from the popular press through policy documents and academic journals, is not easy. It requires time and competence. Those who take a temporary vow of poverty, who have both time and expertise, are a fine source of assistance in the task. They are called "graduate students" and are a natural resource for study of the matter.

An option for the future lies with the National Research and Education Network (NREN). This effort to understand how text and data can be electronically digitized and exploited easily is well underway. To the extent that NREN technology can be exploited to identify instances of "study use" or "data use," that is to the good.

4. The use of statistical data and studies is observable through direct observation and through self-report surveys. Corroboration is important. There are a variety of options.

The first obvious option is direct observation of a study's use in a meeting, by insiders or outsiders, in which

"Low-level, persistent use of information can be important. But traces of it are often weak."

indicators or studies are considered. The interested scholar may, as an independent observer, sit in on legislative or administrative meetings, to record what data or studies were considered by the meeting's participants. This tactic is often expensive, however.

An underexploited and less expensive vehicle for learning who used what data is through committees that fall in the ambit of the Federal Advisory Committee Act. Public committee meetings under the act require minutes or transcripts. Any member of the committee or any attendant of a public meeting can be a tracker of the use of data or a study. Anyone who chooses to acquire and read the minutes of the meetings is a potential expert on the use of certain studies by the committee.

A third option presumes that it is fair to ask the principal investigator (PI) of a study whether the study findings were used and by whom and when. PIs may be well informed or not. The well-informed PI should be recognized and exploited; he or she would benefit from both of these actions. The ill-informed PI might be educated by the question. The principal investigator's report may or may not be accurate. To the extent that such self-reports can be corroborated, they should be.

There is good precedent for full-blown surveys of the potential users of information, a fourth option. Recall, for instance, studies undertaken by the National Center for Education Statistics of school district staff members' knowledge about the information resources sponsored by the U.S. Department of Education. Independence of informants is an important but difficult matter.

There is less precedent for a fifth option: formal surveys of principal investigators who have received funds from a foundation such as NSF. The grant applicant who asks for more money is at times prepared to document users, e.g., the General Social Survey. But most grant recipients are not equally equipped to provide evidence about the usefulness of their work.

Doing surveys and so forth may help to provide evidence about what study or data set appears to have been useful. Prospective controlled field tests are a sixth option dedicated to understanding what could enhance usefulness of studies. Such controlled tests have been run in the mental health arena to learn, at least, that merely providing information is not enough to encourage change.

5. Peer review of research proposals to science foundations is a fundamental device for deciding whether a proposal warrants funding. More important here, the peer review process is an underexploited method of tracking the use of studies completed earlier by researchers who submit proposals.

Experts who are asked to review a research proposal can take into account the earlier performance of the researcher who submitted the proposal. The experts may consider a variety of indicators of the value of the principal investigator's earlier work. The performance indicators might include the uses to which an earlier NSF-sponsored study or data set, generated by the same or other investigators, were put.

There appears to be no uniform, formal mechanism for this kind of capitalization on external reviewers at NSF or at other foundations. Individual reviewers vary in their interest in the earlier performance of a researcher who submits a proposal. It implies that where "use" of a study or data set is important, the data uses that are identified by a scholar who requests funding to do more data collection are important.

6. The durable civil servant is a fine vehicle for understanding what data set or study has been used.

For instance, both Murray Aborn (NSF) and Howard Rosen (Department of Labor) periodically produced "findings" for their directors, findings that could be used to argue that something happened as a consequence of the agency's investment in research. Foundation program staff who rotate through an agency arguably are not relevant to this task simply because it usually takes time for a study to be used in policy or scientific forums full-time. Charging a civil servant with responsibility to monitor data or study use is a good approach if no other options are available. With access to a phone, proposals, and final reports, this amanuensis can turn out periodic reports on the use of reports. Stake suggested that employing a group whose independence is guaranteed would be an interesting option, and this option is worth considering too. Review panels for research proposals might also be exploited productively in this effort.

To the extent that the culture of the civil service agency is changeable, engaging all career civil servants in the task of understanding which data or studies are used then seems desirable. Those who are capable of communication with both PIs and colleagues, and who wish to do so, are in a position to encourage PI's to attend to the matter. Limited resources and legitimate philosophical antagonism toward such a role for the scientist-civil servant need to be taken into account.

7. Focusing only on the use of data sets or studies is misleading. Data production methods are themselves useful products of a survey or study.

For instance, the NSF's support of the Second International Study of Mathematics and Science resulted in comparative data on mathematics achievement. The thoughtful tracker of the uses of data might reckon that the adoption of higher quality survey methods and testing methods is no less important. And indeed, there appears to have been an improvement in the international studies in that principals have agreed upon definitions, e.g., of 9-year-olds and grades, and methods of sampling that make cross-national comparisons more sensible.

Data on the use of methods may be available through self-reports, through monitoring attempts to augment or piggyback on national surveys, monitoring the adoption of survey data or methods in local surveys, and so on.

The slogan "technology transfer," though trite, is apropos. The methods of measurement of academic achievement, the methods of sampling, and so on that are a product of foundation investments are important. The adoption of these methods is important. It ought to be tracked.

There are good precedents for expecting that new methods of producing data are as important as the data set's implications. Precedents for the adoption of new data collection methods are easy to find. For example, randomized controlled tests of programs in criminal justice are now common partly because of the Minneapolis Domestic Violence Experiment. Randomized clinical trials in medicine have become frequent partly on account of the Salk Vaccine Trials.

"To the extent that the culture of the civil service agency is changeable, engaging all career civil servants in the task of understanding which data or studies are used then seems desirable."

8. To judge from empirical study, and as one might expect, certain variables are related to the use of information. The implication is that further empirical study is warranted and, more important perhaps, that one might statistically impute the use of data sets and studies rather than observe their use directly.

There is good empirical evidence that what matters in assuring that data sets or studies are used includes variables such as the potential users' access to the data or study, the quality of the data, the context and complexity of use, and the background of the potential or actual data user.

Each of these variables is in some sense observable. In the absence of any opportunity to directly observe data use, one might impute the use of data from observations on such variables and a simple statistical model that relates the outcome variable—use—to these variables. There appears to have been no published work on such an effort.

9. Policy, strategy, and systems for data use enhancement are important and warrant special study.

Sponsors of studies have helped to enhance the likelihood that a policy-relevant data set or study will be used, notably by investing funds in dissemination, e.g., Rockefeller Foundation's investment in underclass research. Sponsors have been less sensitive to assuring that the effect of this investment is discernible. The Rockefeller Foundation is an exception in that it has asked for independent review of its investment in both policy research and the dissemination of research results.

Data-sharing policy has been adopted by NSF, the National Institute of Justice, the National Heart, Lung, and Blood Institute, and other organizations that sponsor data production and research. This is remarkable relative to many other agencies and foundations. Tracking the sharers is warranted, however. None of the data-sharing agencies have a tracking system, and this invites the invention of a low-cost independent tracking system.

Data enhancement policies and systems that include piggybacking, sample augmentation, and satellite design are promising. For instance, that several agencies cooperate in trying to produce a useful product is worth recognizing. Presumably, all agencies thought the need for the data set or study was sufficiently important to collaborate in the effort to produce it. The collaboration is an easily measured phenomenon and may be taken as an indicator of expected usefulness of a study or data set.

The option of designing multiple independent studies or multiple loosely coupled studies, instead of a single massive study, deserves more attention. The research design issue is whether one ought to sponsor one massive study or sponsor several independent ones if the object is to assure that the resultant data are used. It is certainly easier to manage a big study rather than several smaller ones. But if multiple studies rather than a single study invite more uses then planning muluple studies rather than a single massive study may be productive.

"The research design issue is whether one ought to sponsor one massive study or sponsor several independent ones if the object is to assure that the resultant data are used."

10. When it is important to assure that data or studies are useful in the policy-making process, staying close to the process is crucial. Keeping distant from the policy maker is crucial, too, in the interest of credibility at least.

To the extent that the indicator/study is close to a policy-making process, the closeness can be monitored, for instance, through logs on who spoke to whom, why, and when. Telephone records, speaking records, and so on are vehicles for tracking.

Gaining the distance that is needed to assure credibility, while keeping close, is harder to do. It is not clear how to observe this.

To judge from contemporary empirical research on data use, however, credibility of the source of information that is purported to be useful is important. It is for credibility reasons that some institutions such as the National Center for Education Statistics and the Bureau of Labor Statistics separate the data production function, which ought to be more independent of politics, from the data use function, which ought to depend on the body politic. The General Accounting Office is similarly sensitive to such issues, but meets its concerns in ways that differ from those used at the statistical agencies.

The source of support for a data set or study is also important. To the extent that a sponsor such as NSF or NCES is viewed as dispassionate, the information may be regarded as credible. The public and others do at times register opinions about credibility of sources of information and of sponsors. Formal surveys of credibility of either are possible in principle, but it is not clear how to do this economically.

Searching Near, Far, And Wide: A Plan For Evaluation

Sylvia T. Johnson
Howard University

A Plan for Evaluation

Planning an evaluation for any major national program is a complex task. Often similarities in structure across program implementation in various sites serve as the basis for implementing traditional evaluation designs. If it is a service-oriented national action program, such as Women, Infants, and Children (WIC) or Headstart, there are certain parameters of input, as well as specific outcomes that can be measured and compared, even though specific projects have unique characteristics.

Many funded research programs have common parameters. The requests for proposals may have been structured to elicit examination of certain key constructs, methodologies, instrumentation, or populations, and these may provide the base for evaluation.

Educational research programs of the National Science Foundation (NSF) have goals that are primarily aimed toward expanding the envelope of scientific knowledge and being on the cutting edge of research. Such programs elicit a variety of proposals from researchers with considerably greater variety in terms of constructs, methodologies, and instrumentation than might typically be obtained. They also pose a more formidable challenge to the evaluator.

The Research in Teaching and Learning (RTL) program as well as other divisional programs present delivery models different from traditional school mathematics and science, and projects may vary in size, scope, and focus. Of course, there are intended effects of these programs. However, the variety of approaches and strategies employed, and the broad range of intended effects, spur the search for a method to examine and identify a number of different ways in which these programs may have left their marks—hence, the concept of footprints, left firmly, sufficiently protected from the elements, and molded well enough to be examined, understood, and replicated, and then converted into sturdy trails for the advancement of young learners of science and mathematics.

This paper presents an approach for developing an evaluation of programs composed of diverse projects. A general orientation to the task and the evaluation perspective employed is presented, followed by an overview of the one such diverse program, Research in Teaching and Learning (RTL). That program is then used as an example. Questions that an evaluation should address, and some ways of approaching them, are then presented. In the process of forming the questions, present and former program officers were interviewed. Included are suggestions prepared by a Research in Teaching and Learning Panel convened in the summer of 1992.

> *"Educational research programs of the NSF have goals that are primarily aimed toward expanding the envelope of scientific knowledge and being on the cutting edge of research."*

The Evaluation Perspective

If one could examine a complex program of funded research from an all-knowing perspective, what could be seen? In developing a strategy or plan for evaluating a program of this type, it can be useful to examine it initially from this omniscient perspective, that is, to think of all the things that it would be great to know about it, even though they may be impossible to know—to take an almost "divine" perspective, if you will, and see where it leads you. The broad diversity of research activities funded, especially when that diversity is along so many dimensions—target populations, techniques, methodologies, etc.—further encourages this initial perspective in considering the evaluation task.

"... it can be useful to examine it initially from this omniscient perspective ... to take an almost 'divine' perspective, if you will, and see where it leads you."

This omniscient perspective would go backward and forward in time to examine intention and planning, as well as long- and short-term outcomes. It would cut across all levels of researchers, participants, and other interested or not-so-interested parties. The outcomes would include those conventionally measured, and those virtually immeasurable. It would include the full range of unintended outcomes, both positive and negative, including those unknown and unknowable to the researcher and the ordinary human evaluator.

This perspective would go even further, though, in that it would discern what might have been. The solicitation, review, and selection of research for funding has many decision points, implicit and explicit. Suppose different directions had been taken in the identification of research projects for funding. Would there be important "Footprints" that are not currently in the picture?

Are there areas of desired footprints where we see more evidence of activity? What areas of possible effects show no effects? What footprints are missing?

In a sense, these are questions of ontology. To the logician, the question "what is there?" can be answered "everything," which while true, may not be especially informative, since the elements included may range from the universe to the empty set. Yet they are questions worth raising as a beginning point when the areas of possible effects are broad and diverse. A program officer also noted the need for an epistemological view in determining the extent and value of the "play-off" from funded projects, because the created knowledge is invisible, and the extent of its utilization difficult to identify.

It would seem that this perspective calls for the evaluator to measure the immeasurable, observe the invisible, assess what might have happened if something else had been done, somewhere else, by someone else—a discouraging task, to say the least. In fact, the perspective being advocated here is meant to broaden the sensitivity, thinking, and powers of observation of the evaluator so that a more complete and useful appraisal of the program can be made. When one studies abstract art, or jazz music, or abstract mathematics, one begins to see, or hear, or conjecture more intensely, carefully, and ultimately, more clearly and with greater satisfaction and sense of thoroughness. When one is observing and enjoying a woodland scene, one can see, appreciate, learn, and enjoy even more, albeit somewhat differently, under the guidance of a trained forester, field entomologist, or ornithologist.

The goal of this exercise is to become sufficiently open to experience, information, and ways of knowing so that in developing an evaluation design and examining the many aspects of a complex program one can identify the need to measure a wider range of constructs with more diverse (perhaps, but not always) but less quantitative measures.

As a result, one should begin to see more as one looks more and more carefully, understand the logic of what alternative implementations might have made sense, where they might have occurred, and who might have been the most appropriate persons to have done them.

Crucial to this perspective is an openness and acceptance of alternative ways of knowing (Gordon, 1992), a willingness to question broadly a range of sources, and the time, interest, and wherewithal for sustained observation. Some vital occurrences do not occur often, and only the persistent may receive the reward of witnessing them. Scientific knowledge emerges from careful observation, yet sometimes dependence on conventional documentation limits discovery. While in no way should we expect to discard all of what we know about sound evaluation practice, neither do we limit our observations to conventional models. An approach that is open to receiving data from alternative sources is more scientific, not less so, because it means more careful observation and attending to alternative outcomes (**y**'s from a given **x**, and receptivity to alternate **x**'s as explanations for a given **y**).

This open and questioning attitude means, for starters, the questioning of oneself as evaluator, and repeating this among the evaluation team. It then means that more than the usual suspects are interrogated, and actually listened to.

Conventional methodology, in terms of examining specific projects, describing their inputs, and examining results of outcome measures does have a place in such an approach. In fact, the evaluation could be conceived as having three tiers: the first based on more conventional outcome data from projects; the second focusing on the footprints of the program in terms of impact and utilization, and the third looking for untouched areas, or the absence of footprints. For tiers two and three, the loci of the footprints (or non-footprints) are developed through a series of questions that examine effects on the program, on other research, on practice, and on other institutions.

In the following section, an overview of the RTL program is presented. From the perspective discussed here, a set of possible initial questions is raised. These questions, of course, would be supplemented by others as the thinking continues, and as initial data are collected.

"Scientific knowledge emerges from careful observation, yet sometimes dependence on conventional documentation limits discovery."

Program Overview

The RTL program was begun in 1984 to support new discoveries about how individuals and groups learn, teach, and work more effectively in complex, changing environments. To this end, the program supports basic and applied research on factors that underlie the teaching and learning of mathematics, science, and technology at all levels. The program aims to support cutting-edge research, and has current priorities to look at the following issues.

1. How students learn complex concepts in science and mathematics.

2. How advances in knowledge of mathematical modeling link to learning complex concepts in science.

3. How teachers' subject-matter knowledge and competencies affect student learning.

4. How teachers learn to become inquiring practitioners and active researchers and how they learn to apply that knowledge in their classrooms.

The impact of RTL studies on educational decision making by parents, teachers, administrators, scientists, policy makers, and curriculum developers at all levels regarding student literacy in science, math, and technology of knowledge is an important concern. Program staff also try to incorporate this generated knowledge into teaching methods and educational products that have direct usefulness in educational programs.

The program is aimed at teaching and learning by persons of all ages in formal school settings from elementary school through college, and informal personal and public settings. Accordingly, projects are conducted in broadly differing environments — classrooms, labs, homes, museums, conference halls—with a variety of methods and techniques from the cutting edge of work in these areas. About a quarter of the projects seek to improve understanding of special needs of learners and teachers traditionally underrepresented in scientific careers or whose needs for scientific literacy have not been met. These include women, African Americans, Hispanics, Native Americans, the physically or cognitively disabled, the gifted and talented, and learners whose native language is not English.

Another quarter of the projects examine motivational, attitudinal, or affective factors in learning and teaching with a focus on family, social content, cross-cultural differences, teacher beliefs, or classroom interactions.

The major goal of the RTL program is to generate a knowledge base that informs the national effort to reform mathematics and science education. Within this goal, activities of the program are aimed at achieving the following objectives:

- Supporting research on teaching and learning specific knowledge domains (chemistry, physics, mathematics, biology, computer science, etc.) at both the precollege and college levels, placing strong emphasis on establishing the content and sequence of learning that can be most effective in developing science and mathematics literacy and problem-solving skills.
- Building a coherent and comprehensive base of knowledge on learning and teaching in mathematics, science, and technology to meet future and current needs of decision makers, practitioners, and the research community.
- Encouraging research that will inform the reconceptualization of measures of performance and provide alternative methods for assessing student learning.
- Seeking research projects on the effects and significance of the nature and quality of laboratory experiences at all levels.
- Exploring factors that may influence interest, participation, and achievement in science and mathematics; development of motivation and curiosity; and the making of and persistence in, curricular and career choices at various student ages and educational levels, with a special emphasis on factors that influence underrepresented groups in their choices of course of study.
- Initiating an emphasis on direct teacher involvement in educational research so that questions arising out of classroom practice will

more effectively inform the perspectives, methodologies, and findings of such research.

- Helping assure the application of research findings by teachers, teacher educators, policy-making educational administrators, parents, and other researchers.

What Questions Should Guide This Work, and How Will They Be Answered?

The broad program goal—generating a knowledge base that informs the national effort to reform mathematics and science education—along with the implementation objectives, provides the framework to generate questions. Other questions may be generated by interactions between objectives.

Impact and utilization are clear watchwords of the RTL program. The evaluation design should be centered on these terms, but with two thrusts. The first is a more traditional set of questions, using data conventionally explored in such investigations. These include the following:

- What publications were generated by the study?
- What awards were received by RTL researchers for publications based on RTL projects?
- How many undergraduate and graduate students have been supported by RTL-funded programs? What indices are available on their productivity?
- What conference and seminar presentations have resulted from RTL projects?

The second impact and utilization thrust is a less traditional one, and involves the utility of new knowledge and its effect on practice. Here we are examining impacts from the level of actual classroom practice, through teacher change, to effects on policy formulation in the education and political communities. The impacts of interest are often connected to studies with a rather traditional experimental sort of format, but the evaluation plan should relate to impact of new knowledge on practice. Such a format is the following:

- How do people (children, teachers, etc.) come to know and understand [concept, procedure, or configuration] y? How does [software, metacognition, instructional strategy] x help this process?

The evaluation plan then needs to examine questions of this sort in terms of the entire program.

- What are the influences on classroom practice, in terms of differences in what goes on in the instructional process, and in outcomes for learners? The outcomes should not be confined to problem solving and laboratory skills, although these are certainly of interest. They should include attitudes toward science and mathematics, interest in pursuing a mathematics or science career, interest in electives in science and math, and math and science interest and inquiry orientation, such as use of evidence in decision making, visiting science exhibits and museums, reading popular science periodicals, etc.
- What effects have RTL projects had on the research and develop-

ment community in terms of changes or developments in text materials, computer software, and teacher education? Curriculum material could be surveyed for answers here.

- What has been the effect of the emphasis on videotape technology in RTL projects? Has it had an effect on teaching practice? Are there steps needed to broaden the effects? What are specific examples of high impact? Can these be broadened? Richness of evidence of instructional value and quality is often applied to videotape. What evidence supports this, and does it show impact in practice?
- A reported impact on teachers has been that the research on children's thinking and mathematical understanding has empowered teachers; that is, as they have found that children have "incredible ideas," significant teacher enhancement has been reported. What documentation supports these incidental teacher effects, from studies which actually focus on children's thinking? What techniques would make this effect more broadly experienced?

"The scope of 'non-footprints' or at least fewer and fading ones, is an area of concern."

There is an issue of what is not being done, or is being done insufficiently. The scope of "non-footprints" or at least fewer and fading ones, is an area of concern. Staff have indicated a need to get new players into the research community, and have pointed out the problem of the aging academic cadre. Many research settings are not where problems in our schools are located. Think tanks opt for less harsh surroundings, as do most universities. But should RTL focus more on a broader base of populations? The Eisenhower Project of The Department of Education is important in expanding this direction, but there is certainly need and room for more. Yet there is still the need for basic research, and RTL is one of the few sources funding this work.

- What is the evidence of impact on utilization of new knowledge on mathematics and science teaching on what is actually going on in classrooms, as well as student outcomes in low-income communities, particularly those in schools serving African American, Hispanic, and Native American children?
- The program overview indicates that one-fourth of all projects were aimed at these students. Did these studies involve sufficient resources to maximize impact?
- Are program solicitations distributed to institutions that would be likely to carry out RTL work in inner city settings? Are workshops and professional group information sessions provided to encourage participation?
- What outreach activities related specifically to RTL studies are directed toward newer and nontraditional professionals? To what extent are they involved in panels and related activities?
- Some research centers have been very successful in RTL projects. They have been consistently funded, and their work has resulted in extensive publications, research-related projects, and the development of young scholars. What factors are related to the success of these projects? In what ways can their impact be broadened?

Collaboration is an important objective of RTL projects. It is encouraged within individual research projects as well as across the program. It is cited by staff as a primary objective of all projects. Another objective involves teachers as researchers, both to develop their inquiry and teaching skills and to impact students.

- Have collaboration and involvement of teachers as researchers been used extensively in inner city schools in RTL research projects?
- Has collaboration encouraged new researchers to seek RTL funding?
- How do teachers who have participated in RTL programs feel about collaboration?
- How has collaboration encouraged activity within the scientific community and between the science and math communities?

Here the work of recent years on standards in math and science, and the importance of these for assessments and teaching should be stressed.

In general, the questions above relate to effects on practice, the profession, the development of new research, and other institutions. Several sources of data are implied directly from the questions.

Other types of evidence and methods of obtaining them are found in the report of a 1992 Research in Teaching and Learning Panel. The panel suggested that RTL go back to the planning for the development of the RTL program that occurred in 1977-78, and engage in the following activities:

- Look at how RTL-funded research has influenced research reported at professional meetings.
- Have an independent group evaluate the quality of reviews, both supported and nonsupported, and how the proposers have reacted to them.
- Develop a genealogy to assess the impact of NSF-funded projects on people, i.e., the number of researchers whose initial work emerged out of working on NSF projects as undergraduates, graduate students, postdocs, consultants, etc., and how they developed as professionals.
- Assess the number of people recruited to the field as a result of NSF-funded projects.
- Document the impact of the program by asking people about their impetus into research in teaching and learning (autobiographies).
- Look at comprehensive reports that have reviewed projects funded by RTL.
- Assess the number and quality of journals that have been created as a consequence of the program.
- Assess the research agendas and their outcomes that have emerged from NSF-sponsored conferences.
- Look at PLATO, which has been a hothouse for future developments.
- Provide a snapshot of the people who have served on RTL panels.

"It should be noted that the time and resource base available to the evaluator is an essential consideration."

- Look at research reports from the American Educational Research Association (AERA), National Council of Teachers of Mathematics (NCTM), etc., to assess the number or percentage emerging from NSF funding.

- Look at mathematics and science educators who have broadened their views as a result of interactions with people outside the field, that is, look at the people who have served as consultants and on teams of the projects.

- Assess the time and efficiency of the program relative to NSF structure.

- Do a contrasting analysis of the mathematics and science communities.

- Look at applied journals for both authorship and citations, e.g., Physics Teacher, Science Teacher, and Mathematics Teacher.

- Assess the movement of people into other areas.

- Assess how many proposals in Teacher Preparation and Teacher Enhancement programs and the Instructional Materials Development program build on RTL-sponsored research.

- Assess the extent to which research is blended with practice.

- Look at the research discussed at NCTM conferences.

- Look at how RTL has affected programs at other foundations.

- Assess the impact of research on frameworks and standards.

- Conduct an ERIC keyword search.

- Look at all the regional laboratories and assess what they are disseminating.

- Look at the impact of the research and teaching methods that have been developed as a result of RTL-funded projects.

The questions and data collection sources and procedures above provide a beginning framework for the examination of RTL projects and the impact and utilization of new learning and discoveries. They provide multiple ways of knowing more about the program and its consequences. A parallel examination should provide similarly for other diverse programs of funded research.

It should be noted that the time and resource base available to the evaluator are essential considerations. The evaluator is not in sole control of the evaluation. The approach advocated here requires that the funding source allows for sufficient resources, time, and access to allow the kinds of things to happen that enrich the quality of the data and the evaluation report. If external constraints do not allow for this activity, the evaluation may be severely limited, despite all the openness in attitude conceivable.

Finally, the fact that an evaluation developed using these guidelines focuses on a broader range of evidence than is often considered should in no way be interpreted as minimizing the importance of rigor. Nontraditional does not mean sloppy, nor does it provide an exception to careful, intensive work. In fact, doing such work well is often more difficult and time consuming than working with "hard" data. Rhetoric is no substitute for data, but good science means careful observation and the accumulation of evidence from different sources, carefully and responsibly reported. Nor should a nontraditional label serve as a rationalizing shield for those using traditional statistics poorly, and claiming that their work is not accepted because they "aren't hung up on a lot of statistics." Nontraditional evaluation does not depend on magic: just on science thoughtfully conceived, coherently organized, and clearly reported.

References

Frechtling, J. 1993. *Research in teaching and learning* (background paper).

Quine, W.V.O. 1963. *From a logical point of view.* New York: Harper and Row.

Stake, R. 1975. *Evaluating the arts in education.* Columbus, Ohio: Charles E. Merrill.

Worthen, B.R., and Sanders, J.R. 1987. *Educational evaluation.* New York: Longman.

Gordon, G.W., and Bhattachanyya, M. 1992. Human diversity, cultural hegemony, and the integrity of the canon. In *Journal of Negro education*, 61: 405-418.

New Methods For Evaluating Programs In NSF's Division Of Research, Evaluation, And Dissemination

---

Robert K. Yin
COSMOS Corporation

Basic Nature of Grant Programs and Purpose of This Paper

The National Science Foundation (NSF) sponsors many programs in science, engineering, and mathematics (SEM) education. All of these programs are "extramural," in that NSF makes awards to some performing organization—generally a university or nonprofit organization. The award is usually a grant award, administered under conditions specified in Grants for Research and Education in Science and Education (NSF, 90-77, October 1992). (The programs also make contract awards and enter into cooperative agreements, but these are a very low proportion of all awards and are not the subject of this paper.)

With a grant award, the performing or grantee organization is supposed to conduct a "project." These funded projects become the collective entity known then as the "program," and individual NSF programs routinely issue reports on the nature of these funded projects. (In many circumstances, the work done under these funded projects may not be readily delineated from work supported by other funded projects simultaneously received by the grantee, but this topic also is beyond the scope of this paper.)

The challenge addressed by the present paper is to develop better methodologies for evaluating programs consisting of this sort of infrastructure. Three NSF programs in particular were used as background information for this challenge:

- **Applications of Advanced Technologies Program** ("AAT" program);
- **Policy-Related Research: Studies Program** ("Studies" program); and
- **Policy-Related Research: Education Indicators** ("Indicators" program).

The paper only aims at developing preliminary ideas in this methodological direction and is not intended to be a complete prescription or even operational set of guidelines for carrying out an evaluation. Rather, the goal is to describe why such new methodologies are needed, and then to point to the further methodological work to be done that will lead to the creation of these better methodologies.

> *"The need for new methods derives from the potential inappropriateness of the standard program evaluation model as it might be applied to a grant program."*

Potential Conflicts Between Grant Programs and "Standard" Program Evaluation Methods

The Standard Program Evaluation Model

The need for new methods derives from the potential inappropriateness of the standard program evaluation model as it might be applied to a grant program. Exhibit 1 contains a simplified version of the standard evaluation model. The model puts heavy emphasis on the iden-

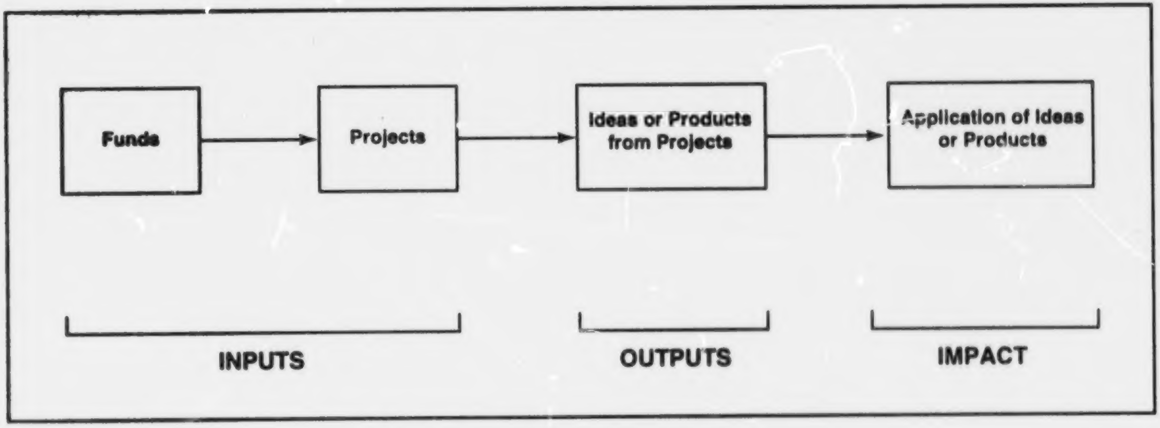

Funds
Projects
Ideas or Products from Projects
Application of Ideas or Products
INPUTS
OUTPUTS
IMPACT

Exhibit 1 "Standard" Program Evaluation Model

tification of real-world impacts. In the SEM education field, such impacts would be expected to occur in actual school systems (K-12 or universities), involving actual teachers and classrooms, and therefore involving actual students. Traditionally, the model also puts heavy emphasis on defining the impacts in quantitative terms. Ideally, the model would like to help policy makers understand how many classrooms and students or teachers were impacted, and to what quantifiable degree, by investing in a particular NSF program.

Attempts to implement the model usually begin with data being collected about the individual projects. The projects may have led directly to applications in the field—and hence may have produced impacts that can be measured. However, if the projects only produce new ideas that are not carried into the field, the model may not be useful. Similarly, the user of standard evaluation data collection methods will encounter difficulties if the impact in the field: a) occurs over a long period of time (say, 10 years) after the ideas were first produced by the project—a commonplace time lag in SEM education; or b) is difficult to attribute because of the relatively small size of the NSF program investment—also a commonplace occurrence because NSF's investment may be in the millions of dollars, whereas the education system of the United States operates at the level of tens of billions of dollars. In either situation, the resulting impacts may be considered overdetermined, and attributing them to NSF-funded projects is hazardous at best.

As a general rule, because education is largely a state or local matter (grades K-12) or a university matter (postsecondary), Federal initiatives must be relegated to extremely minor roles. For instance, the Studies program lists as its major goal the strengthening of SEM education in the United States. Such an impact is very hard to trace, however, given that the program operates with an annual budget of less than $5 million. Similarly, of the three programs, the largest is the AAT program, which supports $10-20 million of funded projects annually in an educational technology market worth at least hundreds of millions (if not billions) of dollars.

At a programmatic level, the interpretation of the results of a standard evaluation also may be little more than the aggregate of all of the project-level results. Strategic considerations pursued by programs—e.g., to overinvest in certain areas of high priority, or to make a few high-risk awards, or to follow any portfolio criteria—tend not to be covered well by the standard evaluation model, as traditionally practiced.

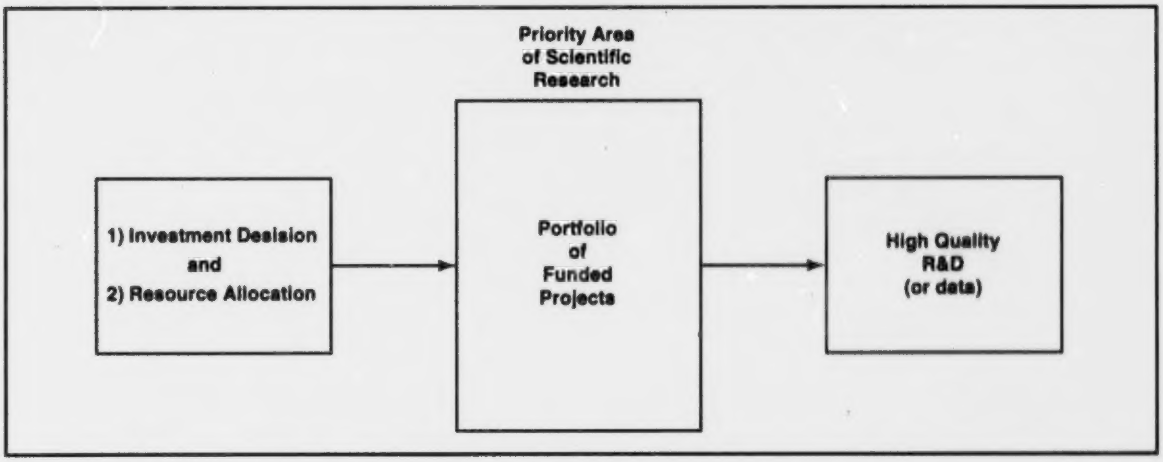

Priority Area of Scientific Research
1) Investment Desision and 2) Resource Allocation
Portfolio of Funded Projects
High Quality R&D (or data)

Exhibit 2 A Grant Program Model

A Grant Program Model

The standard evaluation model does not reflect well the way that Federal grant programs are created, or how the staff or sponsors of grant programs usually strategize about their programs. Exhibit 2 contains a simplified version of how the program might be conceptualized by its staff or sponsors, using a grant program model. Essentially, a public commitment has been made to support R&D in a pre-identified priority area of scientific research. The role of NSF, as a sponsoring agency, is to make these awards in as rigorous and utilitarian a fashion as possible. The main subsequent concern is whether the research was completed in a high-quality manner.

In most grant programs, the grants are used to support basic research. But even where applied research is the main subject of a program, this same type of thinking has traditionally been followed for two main reasons. First, the mandating legislation may contain no specific goals (for example, none of the three NSF programs have specific legislative mandates), and none may have been articulated beyond the statement of need for investing in the area. Second, the award characteristics of a grant mitigate against other ways of thinking. Grant awards deliberately permit grantees to make reasonable adjustments in a project as it starts up and is implemented. Indeed, the purpose of a grant is not to limit an investigation to a rigid design, but to encourage the investigator to make the best choices leading to high-quality R&D. Further, the grant award is considered important in attracting proposals from highly capable investors, who have traditionally been able to take advantage of the independence of grant award conditions to create inventive results.

In the grant program model, the notion of quality would include such criteria as: 1) advancing the state of understanding about a topic ("making a contribution"), 2) developing a framework or foundation for further research on a topic, and 3) far exceeding the standards of a field or academic discipline. Quality may not necessarily include such criteria as relevance to immediate problems, much less having an impact on them.

When a grantee fails to produce high-quality R&D, the major consequence is that—in the long run—the grantee will find it increasingly difficult to obtain new grants. However, other than strictures regarding fraud, waste, and abuse, it is not incumbent for the grantee to "perform" productively on any given grant award. On the contrary, the underlying philosophy is that much new research will fail, and that the nature of research involves a high incidence of failure. In fact, the grant mechanism was designed in part to accommodate this aspect of the scientific enterprise.

"... some scholars reduce the likelihood of failure by performing new studies that are one step ahead of their awards."

Competitive scholars, of course, will always find a way to produce a gain from every funded project. A minor publication, a new descriptive understanding, or a methodological lesson may have to compensate for the failure to complete the original project as proposed. As another variation, some scholars reduce the likelihood of failure by performing new studies that are one step ahead of their awards. Their new proposals therefore contain proposed inquiries whose outcomes are already known, though not yet published or shared with colleagues—and therefore increase the probability of getting a grant award.

The grant model, however, clashes with the traditional evaluation model. The grant model gives little attention to impact. At the same time, high value is placed on quality—which in turn is generally ignored by the quantitative orientation of the traditional evaluation model. In addition, unlike the traditional evaluation model, the grant model highlights the portfolio of projects and incorporates strategic investment goals that are not just the aggregate of all individual projects. For instance, the AAT program prides itself in being a "high-risk, high-gain" effort. In other words, the hope of the program administrators is that a few of their projects will produce scientific breakthroughs, even though the majority of the projects may not lead to significant advances in knowledge. The grant model accommodates this strategic objective more readily than the traditional evaluation model.

Why Evaluation Is Needed

Public investments in grant programs, whether in support of basic research, applied research, or R&D more generally, necessitate the assessment of external benchmarks of progress. Most commonly, the evaluation of a grant program is put into the hands of an expert panel, which may be organized as a "visiting committee" or operate under some prestigious sponsorship such as the National Academy of Sciences. NSF-SEM education programs have been subjected to these types of evaluations as well as numerous other administrative reviews. The challenge is not to displace these efforts, but to ascertain whether formal evaluation methods can complement them.

A New Evaluation Strategy

Formal evaluations can, in fact, be complementary, if only the methods used to conduct them are modified. The modifications are needed to make evaluation applicable to situations such as these occurring in the R&D grant program, in which:

- The intervention is weak or small, relative to the measurable impact of interest;
- The intervention is not a part of a formal research design, because the intervention was not designed to suit the needs of evaluation, but rather to suit policy-related or real-life needs; and
- Extensive time (five years or more) or resources (millions of dollars) are not available to support the needed evaluation effort.

To deal with these conditions, COSMOS's ongoing research has been developing a new methodological strategy (Yin, 1993; and Yin and Sivilli, 1993). The main feature of this new strategy is that it aims to make multiple, partial comparisons instead of imposing a singular research design in carrying out an evaluation. The new strategy offers the opportunity to collect diverse data and to target multiple inquiries in lieu of an overarching research design. The new strategy and how it modifies the traditional evaluation model appears directly related to the evaluation of R&D grant programs.

Exhibit 3 summarizes the traditional evaluation model and its varieties, also showing the niche filled by the proposed new strategy. Randomized clinical trials ("true" experiments), quasi-experiments, and database analyses have all been used in the past as traditional evaluations. The U.S. General Accounting Office (1992) has developed a meta-analytic approach of synthesizing data from these different strategies. The proposed new strategy presents an alternative—filling the gaps between these strategies.

The exhibit shows that when research investigators have no control over the intervention, and when the interventions are not even designed to suit a research design, the need is for some new strategy more powerful than mere database analyses. The new strategy will make some causal inferences possible, even though these will not be nearly as potent as those in quasi-experiments or clinical trials. However, the new strategy may be more generalizable and less costly than quasi-experiments or clinical trials. The new strategy has six features:

- The use of partial comparisons, based on multiple "partial" designs;
- Designation of each single component of a comprehensive program—rather than the program as a whole—as the main unit of analysis (therefore leading to multiple sets of partial comparisons, if a program had several components);
- Greater emphasis on the use of proximal rather than distal outcomes where interventions are of low strength or "dosage;"
- Explicit assessment of the "process" logic of an intervention;
- Replication across multiple components or programs where objectives are similar; and
- Triangulation about key events by using multiple measures.

"... the new strategy may be more generalizable and less costly than quasi-experiments or clinical trials."

Exhibit 3 Evaluation "Niches"

Characteristics of Evaluation Situation

Evaluation Strategy	Researcher Control over Intervention	Interventions Intended to Suit a Research Design	Casual Interpretability	Generalizability	Cost
RANDOMIZED CLINICAL TRIALS	Yes	Yes	True comparisons and causal inferences possible	Trials limited to narrow client populations	High in cost
QUASI-EXPERIMENTS	No	Yes (except random assignment)	Causal inferences possible	Trials limited to narrow client populations	Moderate in cost
PARTIAL COMPARISONS	No	No	Series of partial comparisons possible, ruling out key rivals	Comparisons cover moderate range of client populations	Moderate in cost
DATABASE ANALYSES	No	No	Poor comparisons and causal inferences possible	Databases cover full range of client populations	Low in cost (data already collected)

Of these six features, the most innovative and important deals with partial comparisons, and the remainder of this paper therefore suggests how this feature might work in evaluating a program like the illustrative three programs of NSF.

Application of the New Strategy

Exhibit 4 lists an illustrative set of partial comparisons. The comparisons are considered partial because none alone provides definitive causal evidence about the outcomes of a program. However, each partial comparison is intended to support a positive inference about the program and its outcomes. Thus, the more partial comparisons that an evaluation can cover (and these partial comparisons go beyond the 18 listed in Exhibit 4), the more compelling the argument can be made that: a) positive results were produced, and b) the program under evaluation produced them. The goal of the new evaluation strategy is therefore to identify and collect data that can satis-

Outcomes-Only Comparisons

1. The program performed better than at earlier time (pre-post).
2. The program performed better than another program (cross-section).
3. The program performed better than broader group of programs (cross-section).
4. The program's performance trend is in desired direction (time series).
5. Outcomes appear faster or better than expected.
6. Outcomes exceed initial goals or objectives.
7. Outcomes exceed established standards.

Process-Only Comparisons

8. The program implemented a new set of activities, not previously conducted.
9. The program improved an existing set of activities.
10. The program staff can describe how the program differs from previous policy or practice.

Causal Interpretation

11 The program staff can provide a compelling explanation for a documentable chain of events.
12 Ditto external observers
13. Ditto a key informant (insider)
14. The pattern of outcomes is uniquely related to the program.
15. The intervention is uniquely related to some infrastructure, in turn related to the outcomes.

Rival Interpretations

11R. The program staff can provide rationale for rejecting explanations:
    - general climate
    - competing programs.
12R. Ditto external observers
13R. Ditto a key informant
14R. Ditto pattern of outcomes
15R. Ditto infrastructure

Policy Analyses

16. Magnitude of positive outcomes far outweighs costs of program.
17. Outcomes achieved for the first time in this program.
18. Outcomes generate support for further desirable action.

Exhibit 4. Illustrative Partial Comparisons

fy as many of these partial comparisons as possible. The strategy provides flexibility because the relevant data for each partial comparison and the instruments needed to collect those data may vary. Further, no singular research design is being relied upon; rather, the final evaluation will consist of multiple, partial designs.

A critical subset of the partial comparisons is the explicit consideration of rival interpretations. Unlike database analyses, the new strategy encourages and accommodates the collection of evidence to test such rivals. The identification and selection of rivals is not easy (McGrath, 1982). However, the more the rivals are shown to be untenable, the greater the credibility that can be given to the target program's effects. To this extent, the new strategy should produce more definitive evidence than database analyses.

For the R&D grant program, the application of this new strategy yields a modified model of the R&D program, shown in Exhibit 5. This model shows that an evaluation can go beyond the grant program model (Exhibit 2) and assess the production of new ideas as a legitimate program outcome. These new ideas would be considered legitimate payoffs from any of the three NSF programs. For instance, the AAT program aims at producing new ideas demonstrating proof of concept, the Studies program aims at policy-relevant ideas; and the Indicators program aims at benchmarks reflecting educational progress. However, the model also falls short of the traditional evaluation model (Exhibit 1) in that it does not attempt to deal with program impacts.

Exhibit 6 shows how the modified model can be augmented to incorporate rival interpretations. Two such rivals are shown, although others might also be relevant. The Rival 1 hypothesis suggests that other funded projects produced the same valued ideas; the Rival 2 hypothesis suggests that other programs would have supported the same funded projects in the absence of the targeted program.

Immediate Needs for Developing the New Strategy

This new evaluation strategy cannot be put into place at the current time. Further evaluation or methodological research is needed to refine the strategy and make it operational. As a result, this paper concludes with recommended methodological steps, and not an actual plan for evaluating a real-life program.

The first recommendation is for the development of "measures" of the key program outcomes—new ideas (for research or for practice), influence on policy decision making, and capacity-building of the performer community (where relevant). Conceptually, any measures of new ideas should represent new concepts and new ways of thinking about a problem or situation. Similarly, influence on policy decision making should represent the incorporation of ideas into new decisions. Finally, capacity-building should represent improved skill levels and performance by appropriately trained personnel. Operationally, new ideas, impact on decision making, and capacity-building have generally been identified through peer review panels, such as committees organized by the National Academy of Sciences. Determining whether alternative measures can be developed is the objective of this first recommendation. For new ideas for applications, for instance, the AAT program's operationalization of "proof of concept" is already a

Exhibit 5. A More Practical Evaluation Paradigm Applied to an R&D Program (I)

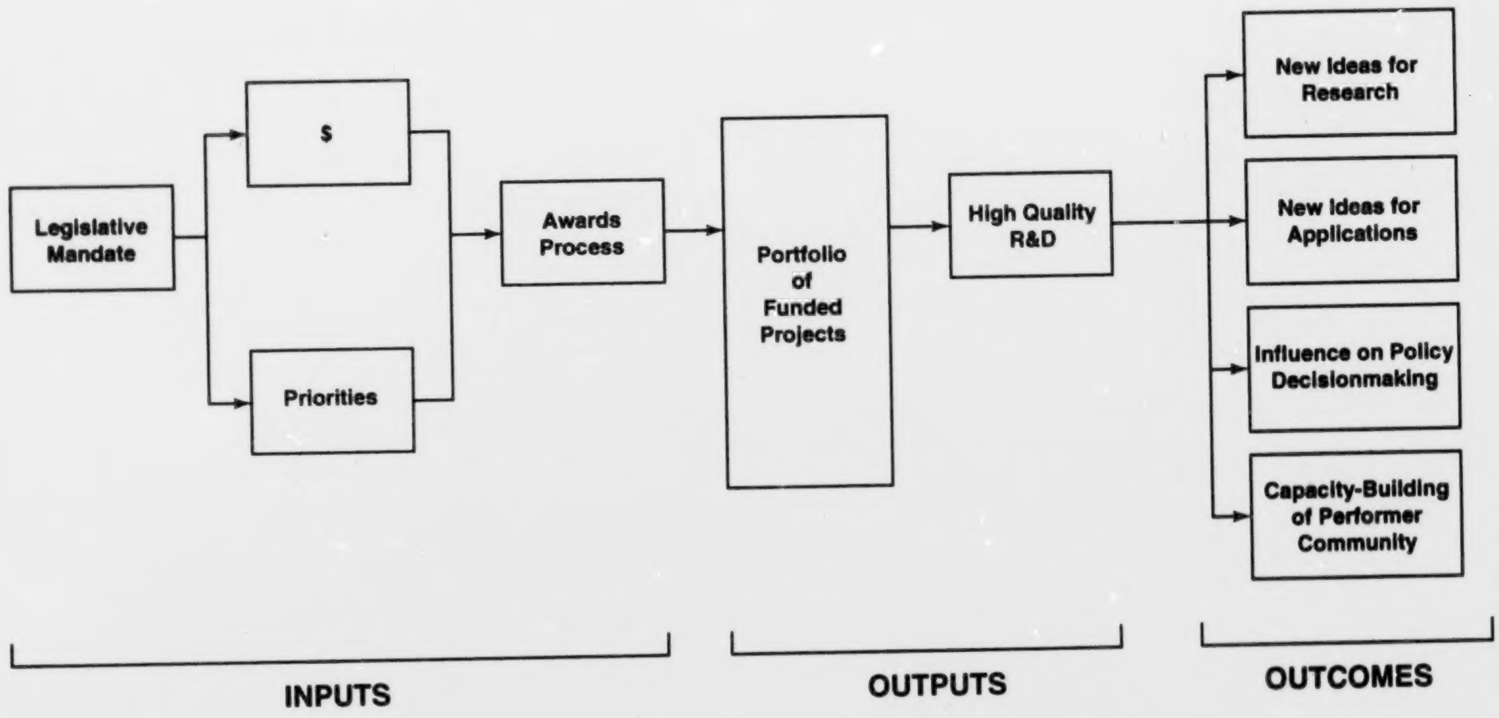
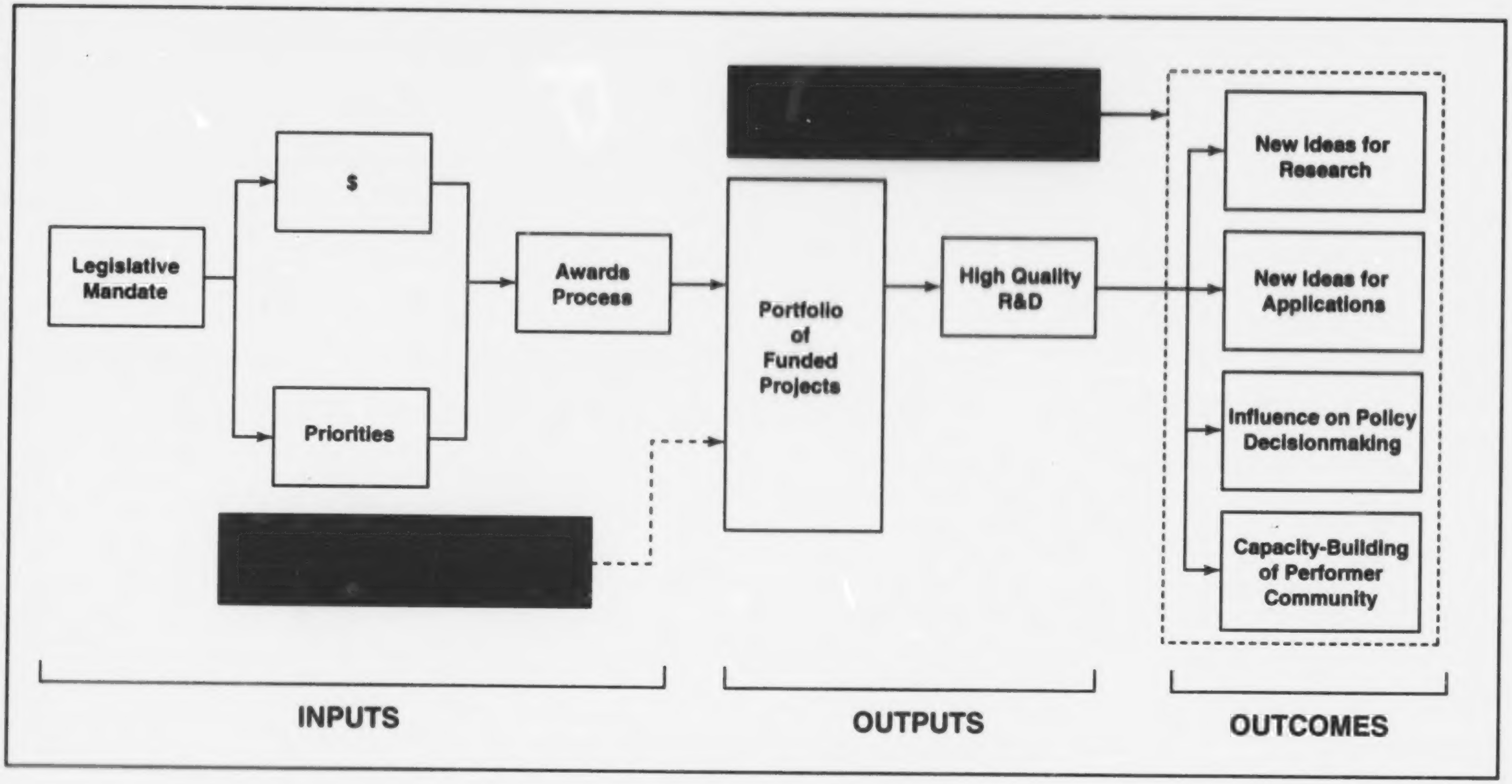

Legislative Mandate
$
Priorities
Awards Process
Portfolio of Funded Projects
High Quality R&D
New Ideas for Research
New Ideas for Applications
Influence on Policy Decisionmaking
Capacity-Building of Performer Community
INPUTS
OUTPUTS
OUTCOMES

Exhibit 6. A More Practical Evaluation Paradigm Applied to an R&D Program (II)


Legislative Mandate
$
Priorities
[redacted]
Awards Process
[redacted]
Portfolio of Funded Projects
High Quality R&D
New Ideas for Research
New Ideas for Applications
Influence on Policy Decisionmaking
Capacity-Building of Performer Community
INPUTS
OUTPUTS
OUTCOMES

promising approach that should be explored further as a methodological advance.

The second recommendation is to develop designs for conducting case studies of funded investigators and the projects they undertake. These investigators may be able to report or demonstrate how they have blended different sources of funds to make different projects or different findings possible. Such patterns might provide clues about the importance of the targeted program, compared to other sources of funds—thereby helping to unravel Rival 2 in the previous example (Exhibit 6).

The third recommendation is to extend the logical list of partial comparisons in Exhibit 4. A comprehensive list is needed, even if any given evaluation can only cover a subset of the list.

Finally, some testing needs to be done to assess the level of effort and costs of undertaking partial comparisons. Exhibit 3 assumed that these costs would be moderate, compared to the costs of conducting randomized clinical trials. However, actual data about the costs would be extremely informative. Some comfort may be derived from an earlier effort (Fitzsimmons, et al., 1992) that managed to track causal program relations within a reasonable time frame and cost limit. This earlier effort did not follow the proposed methodology but did cover a roughly similar scope, evaluating NSF's Coordinated Experimental Research in Computer Sciences (CER) Program.

Summary

The evaluation of ongoing Federal programs—in mathematics and science education and related research—is a challenging problem. The programs already exist, have been operating for some period of time, and were not designed to be part of formal evaluations. An evaluator must therefore address these programs without assuming the ability to manipulate key experimental or treatment conditions.

Traditional evaluation designs do not serve well under these circumstances. As a result, new evaluation strategies are needed. The present paper deals with this challenge by proposing a new strategy of partial comparisons. This new strategy entertains and deliberately seeks to investigate rival explanations and threats to validity. However, the strategy does not assume the creation of a singular evaluation design to deal with all rivals (as do traditional designs). Rather, the total evaluation of a single program will consist of multiple substudies—each potentially using different designs and sources of evidence as relevant.

This paper demonstrates, in a preliminary manner, how the new strategy would be relevant to typical NSF programs in mathematics and science education such as the Applications of Advanced Technologies Program, the Studies Program (policy-related research), and the Education Indicators Program (policy-related research). The paper concludes by identifying the needed methodological work before the strategy can be considered a truly competitive alternative.

"... some testing needs to be done to assess the level of effort and costs of undertaking partial comparisons."

References

Fitzsimmons, S.J., et al. 1992. *An evaluation of NSF's Coordinated Experimental Research in Computer Sciences (CER) Program.* Cambridge, Mass., and Washington, D.C: Abt Associates Inc. and COSMOS Corporation.

McGrath, J.E. 1982. Dilemmatics: The study of research choices and dilemmas. In *Judgment calls in research*, eds. J.E. McGrath, J. Martin, and R.A. Kulka. Beverly Hills, Calif: Sage Publications 69-102.

U.S. General Accounting Office, 1992. *Cross design synthesis: A new strategy for medical effectiveness research.* Washington, D.C: GAO/PEMD-92-18.

Yin, R.K., 1993. Evaluation design: Breaking new ground. Unpublished paper. Washington, D.C: COSMOS Corporation.

Yin, R.K., and Sivilli, J.S. 1993. Evaluation of gang interventions. Paper presented at National Institute of Justice's Fourth Annual Conference on Evaluating Crime and Drug Control Initiatives, June 28-30, Washington, D.C.

Thank you for the opportunity to react to the papers. Coming from an evaluation office charged with producing evaluation reports to inform policy and legislation for the elementary and secondary programs in the Department of Education, I appreciate the clear thinking that has gone into the writing of these papers. The presentation today lifts our sights beyond looking at our day-to-day evaluations in the traditional way.

Our problem in program evaluation studies, and I'm sure this is shared with the National Science Foundation, is that our evaluations are very much tied to the legislative cycles, to budgetary needs, and to looking at administrative changes that have to go on in programs. If they don't do that, they usually don't make it beyond the prospective stage. We rarely have use for studies for which we can't see immediate payoffs.

Further, we must work within some important limitations. Our funding is often dependent on a particular program or a congressional mandate to investigate a particular program. Chapter 1 presents a good example. Because we have a line item for evaluation in the Chapter 1 compensatory education program, it's little wonder that most of the activities in my office concern Chapter 1 and look at issues involving disadvantaged students. At the same time, we need to avoid getting stuck in a rut, relying on boilerplate methodologies when some radical rethinking is really needed. However, currently there is no demonstration authority in the largest of the Department of Education's elementary/secondary programs, Chapter 1. This means that our work is dependent on finding naturally occurring examples of effective practices and programs. Yet we realize that the field desperately needs new approaches to replace the low-level basic skill and drill models that currently prevail. These constraints lead us to take opportunities where we can find them.

Let me share some examples of using opportunities. When sufficient funds were unavailable to launch a full-scale national study looking at math and science programs for gifted and talented students, we scaled back to case studies. These case studies were done by Cosmos, Robert Yin's company. To limit the field of possible sites—we could have gone to hundreds and hundreds—we decided to focus on projects that served disadvantaged students. This resulted in a study that has contributed in several ways to refocusing the Federal effort on assisting the disadvantaged. The study findings were used to craft priorities and selection criteria for both Native American education and the Javits Gifted and Talented program. The study encouraged other work, spurring us to look at strategies from gifted and talented instruction that could be applied to the regular classroom and to examine the impact of these alternatives to conventional wisdom regarding educating disadvantaged students.

We try to stretch our resources and broaden the scope of our evaluations to examine the larger context for Federal programs, rather than always looking program by program. For example, we are currently competing an evaluation contract to examine the Eisenhower Regional Math and Science Consortia and State Curriculum Framework Projects in tandem. It will also look, to the extent we can, at the National Science Foundation's Statewide Systemic Initiative projects. From this study we hope to develop a better understanding of Federal initiatives as they complement or operate independently of each other.

To get more bang from the evaluation buck, we've looked to cooperative efforts across our own evaluation office and with other evaluation offices. Our national evaluation of the Chapter 2 block grant program needed to look at how private school students were participating in Chapter 2—specifically, what special arrangements were being made for their participation. At the same time, we had commissioned a special study to look at Chapter 1, the categorical program, including how private school students were participating. The solution here was very simple. We decided to piggyback the Chapter 2 items on to the larger Chapter 1 study.

Similarly we're working with the Department of Health and Human Services to examine the impact of the JOBS program on the education of the children of JOBS program participants. To study the linkage to adult literacy, we are pulling funding from adult education evaluation funds.

The national performance review initiative by the Vice President has given us a challenge that I hope we can turn into an opportunity. The Department of Education has volunteered to serve as a reinvention lab. It plans to develop performance indicators for our major programs similar to those being mandated in Public Law 103-62. The staff offices in the department, including our own, are also participating. For our part we are developing, with the help of people like Bob Boruch and the members of the National Academy of Public Administration, ways to look at our own productivity and impact. Bob is helping us by developing a user survey similar to the work described today.

I'm thankful to the National Science Foundation for funding the conference and the work of the authors of the papers presented here. Such conceptional work is rarely undertaken without the prospects of immediate payoffs or knowledge of exactly how the work relates to immediate concerns. NSF is making a valuable contribution to evaluation methodology by leaving these footprints. Other agencies can follow them as they go through the process of thinking how to assess the impact of their work and the programs and projects they support.

It's a pleasure to comment on three such intelligent and creative papers. When I first heard of the concept of footprints, it struck me as being of doubtful usefulness, but I've changed my mind.

I like Johnson's paper primarily because she raises both of the two big questions. One question we all ask in this field is, How do you attribute causes from effects? The question we don't ask often enough is, Compared to What? Programs and explanations compete. Johnson longs for the all-knowing perspective, looking backwards to intention and planning, forward to outcomes, and sideways to what might have been. I think some of this sideways vision is possible, as Bob Yin seems to believe. The field of public policy, a major sponsor of program evaluation, does ask very broad questions about what, in a given era, was on the public agenda; what sorts of efforts were deployed (some nonobvious); and what in the end these led to. Although these questions are not very rigorous, eventually there is historical consensus: Were income maintenance plans cost effective? Was the tax cut of 1981 successful? The logical step here is that what might have happened may have happened. It's helpful when, over a decade or more, streams of evaluation are directed so as to flow down ALL the major channels of program and policy reference, not just the main stream. It makes the historical judgment more complete and more sound.

Let me try to relate this to education. Here are three examples of what are essentially competing explanations for certain broad sets of effects. First, in the cognitive realm, there is an established tradition of work in educational psychology that says that the demonstrated level of achievement in knowledge-item testing, at least a variable portion of the score, is a function of the amount and intensity of specific instruction, of actual brain time-on-task in the delivered curriculum. We in Education and Human Resources would not deny this, but we would think the matter more complicated. The point is that this explanation doesn't concern itself with pedagogy or the quality of thinking by the student or the generativity of knowledge: it talks about measured content exposure, the length of the school year, the sequencing of material and the timing of testing, and so on. If the stated criterion is test score improvement, and program evaluation were to show that this molecular and measurable kind of approach yields interventions that pay off, compared to some of what EHR is doing, it might suggest to those who pay us that some of the stones we are lifting are not worth lifting.

Second, at a higher level of generality, there's a "bet" in the nineties that there is a more powerful avenue for the welfare of young people than educational reform: I refer to the well-child movement involving the integration of human services of all kinds, as pioneered at the Harvard School of Public Health and the Carnegie Councils and funded at quite high levels in the Department of Health and Human Services. An implication here is that the dropout rate in high school is perhaps not fundamentally an instructional matter: to explain it, you need to look at the social aversiveness of schooling for some kids, at the labor market and at foregone earnings for these kids, at the family—including nontraditional families—or the neighborhood or the subculture as an economic enterprise, and at still other kinds of explanations. At any rate, this is the kind of situation where in 20 years experts will say which general strategy was "on target"—although if the identified problem has changed, then the desired target may also have changed.

Finally, there is also a bet going that the tocsin sounded in the early eighties about a competent work force, economic competitiveness, and national security is not really something the schools can solve. The argument, now becoming explicit, is, if business needs an up-to-date technically trained work force, let them do the training, invest the capital, and capture the benefits; why load it on the schools?

The general point I am making is that, in a medium-long timespan, if we don't want instances of program evaluation to appear at some later time as quaint or irrelevant, we need to keep in mind the definition of the problem and the public choice arena in which a program existed, compared to other problems and choices. That's why I'm pleased with Yin's Exhibit 6, which begins at the right place and ends ... almost at the right place. Legislative mandate refers, inevitably, to some perceived problem or need, where intervention is thought to be possible. With regard to NSF, the Vannevar Bush report and the 1950 enabling legislation refer to a compact between government and especially the military, industry, and universities that would ensure that a domestic Manhattan Project could be mounted at any time of crisis. Later, in a different era, the report language concerning authorization and appropriations for EHR during its rather extraordinary period of budget expansion gives us various statements about why, for what purposes. The corresponding language for the Department of Education presumably has addressed other large issues: the dropout rate, the school-to-work transition, and the problems of multilingualism and multiculturalism. It is important in program evaluation to examine the sense of problems, needs, and possibilities that existed as the program itself came into existence. All I want Yin to do is to bring that analysis around to the right-hand side of the figure, so that we see outcomes with respect to what. That is, what do the new ideas, applications, capacity, and so on address? Is it leading a good life? Is it economic viability at the personal and societal level? Is it raising achievement in school?

This bears directly on Yin's commendable inclusion in his model of two locations for rival hypotheses: that is, competing explanations. The two boxes represent different sorts of processes. The box at the top, subscript 1, refers either to historical convergence of cause to effect processes or to alternative causes or paths to the same effects. That is, these same ideas, applications, influences, and capacities would occur anyhow, for different reasons. In that case, the program in question was in synch with other cause-to-effect processes; at worst, it duplicated them unnecessarily.

The box at the bottom, subscript 2, refers to a narrower kind of explanation: that normal science, including "normal" applied research, is highly overdetermined, reflects the Zeitgeist and runs under its own steam. It is not genuinely *directed* toward the ends shown in the chart, though they may indeed be true consequences. The challenge is that the specific mandate, appropriations, priorities, and funding decisions of an NSF program contributed nothing distinctive: the availability of any orderly decision process would have led to the same quantity and quality of R&D. Examples: there is a technological shift lying behind Research in Teaching and Learning; there is a particular public policy research agenda driving the Studies program.

We are more familiar with this latter kind of "compared to what" challenge in evaluation of granting programs. I have two specific suggestions. First, it is useful to map the portfolio of funded projects onto the set of all fundable research projects: projects designed, proposed, field tested, or conceptualized by a given pool of researchers. If what NSF selects is basically an exact subset of all possibilities out there, across a defined set of research generators, then there is a tight relationship between the field and the program. The field drives the program, the program fuels the field. This is said to be the case in some programs at NIH, where a successful grant-getting investigator always proposes the research he or she has just successfully piloted (or even completed). If the two distributions are not alike, it may be evidence for a specialized ecology, some sort of lock and key fit in research funding: some proposals go to NSF, some to

ED, some to Spencer, and so on. In this case, the differentiated route to outcomes is more easily traced.

I would like Yin's box, Portfolio of Funded Projects, to be shown in relation to another box, called Portfolio of Possible Projects, in some other plane or orientation. This comparison is not done often enough; it is feasible, but it is difficult. As these papers point out, investigators work on different things under the same grant, or on the same thing under different grants, etcetera. Since the outcomes in question are not always measurable in terms of money, it is impossible to construct their production functions in the usual econometric terms. So my second suggestion is to use *time* as the metric. In principle, it is feasible to go into the population of those doing educational research and ask about investments and yields (appropriately discounted) and opportunity costs. Why did you do this research rather than that? When did you expect a payoff? When did it arrive? How much time have you spent not doing research, but volunteering in a high school classroom? Serving on a school board? Lobbying for specific educational practices at the district office or the state house? Teaching a course in the School of Ed—if you're a departmental scientist—or accepting an education graduate student for a dissertation? Urging young faculty to go out into the schools ...? Johnson, in her paper, suggests some of these possibilities, and there have been some useful studies by the Woods Hole circle around Zacharias and Bruner in the early sixties along these lines. After all, researchers choose among research possibilities, and they are not just researchers. If real impacts and outcomes in the educational arena are to be attributed to a full range of causes, or even if the dynamics of the research process are to be fully understood, then these "compared to what" tracings and paths are important.

I apologize to my esteemed friend Bob Boruch for not delving deeply into his paper in this forum. He knows that I think it's full of good ideas. Briefly, I endorse the importance he gives to filter mechanisms and intermediary groups: these are key aspects of both quality control and uptake of information. Overlooked sources of unique information about knowledge into practice include, besides those Boruch mentions, scholarly autobiographies, Festschrifts documenting intellectual circles and institutional histories (e.g., of the Education School at Stanford), and retrospective why-I-worked-on-what-I-worked-on-when-I-worked-on-it volumes such as the one Rossi did for the Russell Sage Foundation a few years ago. And the idea about tapping into the memories of longtime civil servants can be extended to certain retired agency officers, who can give crucial information and advice at important moments without their egos being on the line. (You remember how in John le Carre novels Smiley was always being brought back from retirement or disgrace, because they needed him at Cambridge Circus.)

One thing Boruch just touches on (as does Yin) but which is very important, is that in the grant-giving arena it is impossible to trace effects to causes if the only information used is what the researcher *proposed* to do. All the agencies and most of the foundations do a poor job in documenting what was actually done. Program evaluators are quite familiar with this problem, but it's time we in the agencies took some of the burden off them by doing a better job of record keeping and documentation of first-level outcomes ourselves, that is, what the intervention or activity actually amounted to.

I represent a mission agency, NASA. We are not the National Science Foundation. We are not the Department of Education. Our programs have a specific kind of very results-oriented approach. We have a mission to carry out, and that determines the kinds of programs that we can do.

I was very pleased to discover that, while all the papers described what were called nontraditional research methodologies, I didn't find them nontraditional at all. They all model what should be, and is, good evaluation practice. They are only nontraditional in the sense that they are not often carried out in Federal government work.

One of the things that came through in several of the papers, and which I think is important, is the unit of analysis that should be looked at in evaluating programs. That is, what is the distinction between a program and a project? People often confuse the two. At NASA, for example, we have over 300 different programs, many of which are, in fact, actually small projects. I think each of the papers, in different kinds of ways, encourages us to look at the impact of these projects in the aggregate rather than as individual small effects. Such small projects are going to have a limited impact in that the effects, if not immeasurable, certainly will not be very useful to anyone.

There is also a lot of discussion about the difference between quantitative and qualitative data. I have reflected on this since I have been in the government. Why do we spend so much time and emphasis on the collection of quantitative data about our programs? (How many teachers were served? How many curriculum products have we turned out? and so forth.) I blame that machine—the overhead projector. I feel a little bit vulnerable here because I am not using viewgraphs, and in the government, as well as many other organizations, there is a point where you have to present information about your program that can be summarized on one or two viewgraphs. That almost requires a quantitative approach, so that you can build a little chart with numbers and statistics. I give this challenge to myself, as well as to my colleagues and to the writers of these research proposals: think about creative ways to present the results of research that uses qualitative data and a variety of very creative analyses of all those data. Think about how qualitative information can be summarized and communicated in an effective way so that it really will have an impact on future program operation.

There was not much discussion about needs assessment in the papers, and I think it is very important for all of us evaluators to pay closer attention to that issue. There is a recognition that what drives programs in the Federal government is legislative authorization. But, in many cases, there is a great deal of flexibility. There are options, different choices that can be made about the programs. Those options should be selected on the basis of comprehensive needs assessment, which is almost never done.

Finally, I would like to thank Dr. Boruch for teaching me a new word in his paper, "amanuensis." I was not familiar with that word. For those of you who don't know what it is, it is someone who writes from dictation or copies manuscripts. Very often I feel like this at work, and I think many of my tired colleagues feel the same way. Maybe if we expand our horizons in the production of evaluations our vision will be brightened and our work will become more creative and meaningful.

Considerations For The Evaluation Of The National Science Foundation Programs

Richard T. Hezel, Hezel Associates
Syracuse, New York

Introduction

This paper contains a set of considerations and suggestions for the evaluation of any National Science Foundation program. Of special interest to the writer is the Applications of Advanced Technologies (AAT) Program, which seeks to generate knowledge on the applications of new advanced technologies to the learning and teaching of mathematics and science. Moreover, the AAT Program strives to inform researchers, policy makers, decision makers, vendors, and developers of instructional materials about the research associated with funded projects.

An initiative that focuses on rapidly transforming technologies, the AAT Program, by its charter and mission, requires flexibility. The program accepts certain inherent risks in the funding of advanced technology projects, some of which may meet outstanding "success," while other funded projects may appear to "fail."

Program Profile

The AAT Program has supported projects whose goal is to investigate the development and use of advanced technologies, as well as projects that permit the broadest dissemination of information about the uses of technologies in various settings. AAT has supported research on, and uses of, innovative, cutting-edge technologies that have not previously been applied to particular uses in math and science education. Because the program supports advanced technologies, the program's goals, along with some of the technology applications to be supported, have changed somewhat over the years to reflect concerns with innovative, experimental technologies that might have applications in education.

By most standards of experimentation, "successful" projects yield outcomes which are desired, hypothesized, and expected. In some cases, unexpected outcomes, though not originally desired, generate results that are unforeseen, but still positive. In other cases, while the hoped-for results might not be realized, the project might yield valuable information that has long-term effects on the program and subsequent projects.

By its mission, therefore, AAT tends to support high-risk projects in which a "successful" outcome is uncertain. If successful, the projects also have the potential to provide a high payback to the education community at large. As a result, AAT has been willing to accept a higher risk and a potentially higher "failure" rate for funded projects. For the evaluator, such high risk/high gain outcomes present a challenge of assessing the value of the project outcomes, particularly when a substantial number of projects may not produce the desired results.

Part of the value of the program resides in project grantees' abilities to quickly disseminate information about their findings. Regardless of the project outcomes, the application of new technologies in learning settings requires that

"In some cases, unexpected outcomes, though not originally desired, generate results that are unforeseen, but still positive."

project results reach potential technology users as broadly and rapidly as possible.

Introduction to the Evaluation

Since 1984, NSF's Directorate for Education and Human Resources, through its Applications of Advanced Technologies Program, has funded projects designed to generate knowledge on the applications of new advanced technologies to the learning and teaching of mathematics and science. NSF is currently engaged in planning for the evaluation of the AAT Program, and this paper has been prepared to assist that effort. Specifically, it describes potential approaches to evaluation of the program, methods that might be useful in evaluation, and special considerations for evaluation due to the innovative nature of the program. While each NSF program carries out internal evaluations, primarily through committees of visitors, this evaluation project represents the first attempt at an external evaluation of several NSF programs. As such, regardless of the frequency of evaluation, the current evaluation should be perceived as part of a system of self-renewal (Worthen and Sanders 1987), not as a discrete study oriented toward specific decision outcomes.

Program evaluations may be designed concurrently with program development or added subsequent to the program's development and initial operation. In general, the evaluator's role is more broadly defined where the evaluation is planned during the program's development. In such a case, the evaluation is collaborative with the program, administrators and grant recipients. Then, the evaluation itself is viewed as part of a continuing process in the life of the program, and all participants view the evaluation and its outcomes as central to the program's development.

By virtue of the fact that the current evaluation was conceived after the program had been operating for a considerable time, the question of what to evaluate, how to evaluate it, and what to observe presents a significant challenge. By design, the evaluation is "post hoc," in that many of the observations are made in retrospect.

Ordinarily, in the retrospective approach to evaluation, especially one that follows many years of program operation, valuable data collection and observation opportunities are lost. In particular, the opportunity to collect data that measure progress toward goals is absent. Regardless of whether the evaluation is goals-based, the post hoc evaluator has fewer options in the observation of outcomes of the program and its projects.

A concern that often arises in the evaluation process is the "intrusion" of evaluation in program design. Clearly, in the post hoc evaluation design, the evaluation cannot be said to inhibit the program design, because the program is designed independently of any evaluation plans. Therefore, despite its limitations in data collection during the progress of the project, the post hoc evaluation has substantial merit.

The post hoc evaluator has neither precedents for the design of this evaluation nor historical, systematically collected data that might contribute to it. Therefore, the evaluator is relatively free of predeterminations and biases that might have been introduced by design precedents and historical data schemata.

Approaches to the Evaluation of NSF Programs

In light of the developmental history of AAT and other programs, several alternative evaluation directions are evident. The principal approaches can be labeled broadly as objectivist and subjectivist. While a systematic, objectivist approach may be desirable for the evaluation, it may not work effectively because of the complex phenomena to be observed in such a program and in the projects that the program funds. Therefore, a subjectivist approach, which accounts for a variety of phenomena and various methods of measurement, would seem more appropriate.

To what extent should the evaluation rely on programmatic goals to set the evaluation agenda and scheme? In light of goals established for the programs, some combination of goals-based and goal-free evaluation seems warranted.

Goals-Based Evaluation

In cases where programmatic goals have been clearly established during the program's formation, the goals and the subsequent concrete and precise objectives become the criteria for measuring the "success" of the program. The goals-based approach is particularly useful for evaluating those aspects of the program that are circumscribed by goals established for the program. In this case, the goals established for the program articulate in a general way the outcomes expected from the program. In turn, the expected outcomes form the basis for the measurement of actual outcomes.

The AAT program has some general goals and objectives, which could form the basis of an evaluation. Nevertheless, a goals-based evaluation project, to be successful, requires the important intermediate step of validation of the goals as historically accurate and representative of administrators' intentions. A pre-evaluation paper summarizing the AAT goals is an important step toward a goals-based evaluation. A goals-based evaluation, in which outcomes are compared to goals, is desirable for part, but not all, of the evaluation. It is important to note that the goals-based component of the evaluation is not to be construed as utilizing a discrepancy model. The discrepancy model chiefly seeks differences or discrepancies between goals and outcomes. As a result, the model is "problem-oriented," and therefore biased toward negative evaluations.

Goal-Free Evaluation

In the absence of clearly articulated goals, or where articulated goals do not appear to circumscribe the sum of possible evaluation criteria and data to be collected, a naturalistic approach is appropriate. Such a strategy permits the collection of data from multiple sources in a retrospective manner free of the constraints of goals and their outcome expectations. Based on the description of the program and information gleaned from the clients and stakeholders, the evaluator organizes potential sources and locations of data and collects available existing and new data.

Scriven (1972) would most likely argue primarily for a goal-free evaluation, particularly where either goals are not clearly articulated or where the goals do not delineate the likely outcomes. The goal-free evaluation avoids the narrow focus of pre-established program goals and allows the evaluator to focus on actual outcomes, including unanticipated outcomes, rather than intended program outcomes only. A goal-free

---

"... where programmatic goals have been clearly established during the program's formation, the goals ... become the criteria for measuring the 'success' of the program."

---

evaluation is likely to increase the likelihood that unanticipated side effects, both positive and negative, also will be noted.

"Footprint" Evaluation

A type of goal-free evaluation that is both phenomenological and constructivist has been labeled "Footprint" evaluation. Free from the stringent limitations of traditional, management- or objectives-oriented, goals-based evaluations, the investigator examines the project outcomes not anticipated by goals. Of particular relevance in the NSF evaluation are the short- and long-term effects of the programs on their various stakeholders and nonstakeholders. The outcomes of each funded project can be observed most centrally and efficiently at the level of the project director. The broader outcomes, especially secondary influences of the project, require the evaluator to cast a wider observation net.

The assessment of dissemination efforts and outcomes especially crystallizes the trade-offs that occur in selecting either a goals-based or a goal-free evaluation approach. In favor of the goals-based evaluation, the more planned the dissemination has been, the greater the likelihood that dissemination outcomes will be traceable and identifiable. At least the evaluator has clues about where to look for evidence of dissemination attempts, so that the efforts might be assessed and future footprints will be identifiable and identified.

Dissemination Evaluation as an Example of Footprint Evaluation

The dissemination process raises other issues for the "Footprint" evaluation and provides pertinent examples for goals-based and goal-free evaluations. From the perspective of the goal-free evaluation, the evaluator observes possible dissemination outcomes, somewhat systematically and randomly, but anticipating where they are most likely to occur. The investigator searches in many and various places, not just in the places where the planned dissemination was to occur.

In particular, the effects of project information dissemination may be most effectively assessed in their potentiality, that is, the dissemination efforts attempted that are not part of the actual or real impacts of the project on the profession and the public. As demonstrated in NSF program goals, the dissemination process is vital to program success. Therefore, project dissemination attempts should compose a major portion of the evaluation.

Among the dissemination questions to be treated by the evaluation are the following:

- How and to what extent do project information and outcomes influence the variety of publics who are among the target groups of the program?
- What impact does the project have on individuals in the education profession and other institutions in terms of the development of ideas, research, and practice that emanate from the funded project?
- What new research and applications are undertaken as a direct result of the funded project and its findings?
- To what extent has the funded project yielded information that has been widely disseminated to groups and individuals in education and in business?

While the potential impact assessment of the project and its dissemination are identified above, the actual dissemination process should also be evaluated. Included in the process evaluation are the type, methods, and extent of both planned and unplanned dissemination of project results.

Evaluation Orientations

To be avoided in the evaluation of NSF programs is a utilitarian approach, which would suggest that the value of any program rises in direct relationship to the number of people the program serves successfully (House 1976). Applying such an evaluation scheme to NSF programs and their funded projects would result in the predetermination that projects that serve the most individuals, either directly or indirectly through information dissemination, would have the highest value. While the indirect influence of the programs and projects may be immense, there are limits in the ability to adequately measure the sum of the influence.

As for any program evaluation, the evaluation of NSF programs demands the recognition of one or more orientations or clients. Principal orientations or clients of the NSF evaluation are NSF administrators, consumers or taxpayers, and experts in the fields of investigation.

On behalf of the program management, the evaluator seeks to identify the decisions the administrator might make and collects useful information that demonstrates the advantages and disadvantages of each decision alternative. Program modification and improvement are examples of decision alternatives of the NSF program evaluations. The management-oriented evaluation assumes that the administrators are the clients of the evaluation and that they seek the evaluation findings.

A consumer-oriented approach to the evaluation of programs has the taxpayer-citizen as client. Through a consumer orientation, the evaluator seeks to determine the effectiveness of the program in terms of its value and service to the public, however that public is defined. When combined with a management approach and applied to NSF programs, the evaluation takes more of a public interest stance: How is the program benefitting the public citizens in general or some broad group of individuals the program is intended to serve?

Because of the esoteric nature of some NSF programs, there is considerable value in an expert-oriented evaluation, which relies primarily on the subjective professional judgment of experts in the fields of research whose outcomes are being evaluated.

The clients represented above can be considered stakeholders in the NSF programs. While a stakeholder evaluation alone, as such, is not advocated here because it lacks necessary breadth, it is important for the evaluator to consider the client/stakeholders as both sources of data and as groups to observe for the collection of data. Among the stakeholders are

1. The funders, NSF administrators, and program administrators;
2. The grant recipients and their associates who execute the projects;
3. Direct recipients of the project results or information dissemination, mostly in the academic and technology business communities;

> *"To be avoided in the evaluation of NSF programs is a utilitarian approach ..."*

4. Indirect beneficiaries of the project results, mostly the public at large; and

5. Possible unintended "victims" of the program, such as taxpayers, groups systematically excluded from projects or their outcomes, and people who suffer negative side effects of otherwise useful projects.

Data Collection for the Evaluation

The measures of the goals-based evaluation flow directly from the operationalization of the goals, and tend to be more quantitative than qualitative. The Footprint evaluation requires a different set of data collection and methods from the goals-based evaluation. Data are collected more "naturalistically," with an emphasis on qualitative, as opposed to quantitative, data. Some of the measures, methods, and evaluation targets are described below.

Recommended Evaluation Topics and Measures

- Assess perceptions of the project, especially project outcomes, through interviews with project directors, their colleagues and associates, participants, and other experts who are familiar with the field.

- Assess the number and perceived value of new ideas and models of learning and teaching with technologies created and tested under NSF sponsorship or stimulation.

- Assess experts' perceptions of the ideas and models created and tested under NSF sponsorship.

- Assess experts' perceptions of funded projects and the value of project outcomes.

- Assess experts' retrospective and current perceived value of NSF-supported research and development on applications of advanced technologies, especially with regard to innovativeness, national impact, and uses of advance technology for learning, thinking, and problem solving.

- Assess the perceived "usefulness" and value of research on cutting-edge technology.

- Estimate the extent of uses of program-supported advanced science and mathematical concepts by educational leaders and in classrooms.

- Estimate the capacity of students to cope with problems of increasing abstraction and complexity at earlier ages.

- Analyze the results of pilot testing of new concepts and prototype materials in schools and colleges, especially with regard to the understanding of how and when new ideas can be introduced into the curriculum.

- Analyze the results of dissemination of all research completed under NSF support, including scholarly articles, articles in professional publications, news coverage, and presentation in other media.

- Undertake studies of public awareness of key concepts developed and disseminated under NSF support.

- Assess how and to what extent project information and outcomes influence the variety of publics who are among the target groups of the program.

- Assess the impact of projects on individuals in the education profession and other institutions in terms of the development of ideas, research, and practice that emanate from the funded projects.

- Determine what new research and applications are undertaken as a direct result of the funded project and its findings.

- Assess the extent to which funded projects have yielded information that has been widely disseminated to groups and individuals in education and in business. Identify the type, methods, and extent of planned dissemination of project results, and the type, methods, and extent of unplanned dissemination of project results.

- Assess the number of minority individuals participating in a program; the type, number, and effectiveness of minority outreach efforts; and the number of minority groups and individuals reached.

- Assess the program's impact on teaching and learning among individuals who have participated in the project and among individuals who have been reached by the program dissemination efforts.

- Assess among grant recipients the sources and origins of project ideas and goals, including the role of NSF funding and support in the generation of the ideas.

- Investigate follow-up activities to the grant activities, in particular what new research, projects, and dissemination have occurred.

- Track the planning of future anticipated directions and applications of the funded activities.

- Determine from principal investigators the duration of projects and the difference between the proposed and actual duration of each project.

- Assess investigators' initial goals for research and project activities, unanticipated findings that emerged from the research, and other research that has been pursued outside the scope of the grant or the project plan.

- Assess the effects of the project on participants, their attitudes, and their learning, and perceptions of the role of the project in their lives.

- Assess the impact of the projects and their activities on the professional activities of other individuals and organizations who have used the projects and their findings for other purposes.

- Conduct a thorough document analysis, including a review of each proposal and final report to determine initial goals and actual outcomes. Conduct interviews of NSF program decision makers regarding feedback received from past recipients, how past-funded project results affect future funding goals and decisions, and how the project results guide the formation of future goals.

- Assess criteria NSF uses to determine the "success" of projects and how NSF decision makers arrive at the criteria.

- Assess the methods NSF programs use to decide which projects are to be funded. Determine what predictors of success are applied from past projects.

Conclusion

The paper has offered recommendations for the evaluation of NSF programs. Given the posthoc nature of the evaluation design and the absence of identified measurement criteria, an evaluation that is solely goals-based carries serious limitations and is ruled out. Instead, a combination of evaluation approaches that includes both goals-based and goal-free methods is necessary and recommended. In general, the goals-based approach is seen to be valuable in the measurement of anticipated project outcomes, while the goal-free approach assesses broad effect, including unanticipated effects. Both approaches utilize quantitative and qualitative data.

References and Citations

Guba, E.B., and Lincoln, Y.S. 1981. *Effective evaluation.* San Francisco: Jossey-Bass.

House, E.R. 1976. Justice in evaluation. Vol. 1. In *Evaluation studies review annual*, ed. G. V. Glass. Beverly Hills, CA: Sage.

House, E.R. 1983. Assumptions underlying evaluation models. In *Evaluation models: Viewpoints on educational and human services evaluation,* eds. G.F. Madaus, M. Scriven, and D.L. Stufflebeam. Boston, Mass: Kluwer-Nijhoff.

Scriven, M. 1972. Pros and cons about goal-free evaluation. *Evaluation Comment* 3(4): 1-7.

Talmage, H. 1982. Evaluation of Programs. In *Encyclopedia of educational research*, 5th ed. H.E. Mitzel, New York: The Free Press.

Worthen, B.R., and Sanders, J.R. 1987. *Educational evaluation: Alternative approaches and practical guidelines.* New York: Longman.

Communicating The Value Of The National Science Foundation's Contributions To Research And Innovative Technical Applications For Mathematics And Science Education

Norman L. Webb
Wisconsin Center for Education Research
University of Wisconsin

Introduction

When evaluating National Science Foundation (NSF) programs that fund research and innovative technical applications in mathematics and science education, it is important to consider the main purposes of the evaluation. One well-established purpose calls for the evaluation to identify the effects of the program—on the profession, on other research, on practice, and on other institutions. But focusing attention on effects tends to direct attention away from the intended audience of the evaluation. Stated differently, the evaluation of NSF programs should not only identify the effects of programs, but the evaluation should communicate the value of those effects to a variety of audiences—the United States Congress, the mathematics and science education professions, the NSF administration, and the public. Indeed, the value of research and innovative technical applications is often greater than its immediate effect, and any evaluation that fails to communicate this value will fail to live up to its potential. Consequently, in the effort to design nontraditional approaches to evaluation that is presented in this paper, I argue that the determination of a program's value should be integrated with the communication of its value.

A full appreciation of promising approaches to the evaluation of NSF programs that fund research and innovation requires an understanding of several factors. The Research in Teaching and Learning Program, the Applications of Advanced Technologies Program, the Educational Indicators and Studies Program, and other NSF programs are complex. Each program has multiple goals, incorporates expectations that are not always clearly articulated, uses limited resources to solve large problems, and is required to be sufficiently flexible that it both responds to immediate concerns and prepares the Foundation for future needs. Given this complexity, a productive evaluation of these programs should draw upon knowledge of at least the following:

- The nature of research, innovative development, and research-driven enterprise;
- The long-term pay-offs of some kinds of research;
- The most promising lines of inquiry at any given time;
- The past record of established researchers;
- The need to nurture young researchers; and
- The relative importance of groups that have a special interest in the research.

The evaluation of NSF programs that fund research and innovation is further complicated by the nature of mathematics and science education. The teaching and the learning of mathematics and science are different. Each field has

> *"... the evaluation of NSF programs should not only identify the effects of programs, but the evaluation should communicate the value of those effects to a variety of audiences ..."*

different curriculum needs and traditions. Those who work in or interact with each field vary greatly in their interests, work, and demands that are placed on them. This observation applies with equal force to teachers, teacher educators, students, researchers, scientists, mathematicians, school administrators, and policy makers. Each field of mathematics and science education has its own community of scholars and researchers. Nonetheless, NSF programs must serve both fields and, at times, must even allocate resources among the researchers who work in both fields.

As a body of inquiry, evaluation itself adds to the complexity of determining the value of governmental research programs. Studying and evaluating an NSF program has political overtones and ramifications. In addition, amidst calls for public accountability for programs of this kind, the task of assigning a value to the work of the program may create some troubling paradoxes. Specifically, the evaluation of research that is carried out under a given program may validate the high quality of one set of research findings that run counter to the findings of other well-publicized and developed projects supported by the same agency. Further, because each NSF program funds a wide spectrum of projects, this situation could even occur within an individual program. Finally, the costs of evaluation also add to the complexity of determining the value of programs. The benefits of an evaluation to the program and the Foundation must be weighed against the expenses of conducting evaluation that can adequately deal with the multifaceted composition of the program. Since these kinds of factors are important practical constraints upon program evaluation, they should be considered when designing and selecting models for the evaluation of NSF programs.

In this paper I have tried to speak to some of these concerns. However, a full explication of these factors and their relationships to evaluation would require a major document. My intent here is to offer sufficient explanation that the rationale for each nontraditional approach to evaluation is made clear.

In brief, I recommend a series of evaluation studies, since the varied characteristics of the studies best accommodate the variety of goals that a program can have.

- One recommended study, a retrograde analysis, considers how funded projects have built on and used findings from previous projects that were funded by a program. The retrograde analysis is designed to communicate the integrity of the programs and to show how funded research and projects have built on each other to develop a body of knowledge that is being applied to science and mathematics education.

- A second proposed study, a video documentary, is designed to use visual images to communicate the findings and innovations that have been generated through NSF programs and to elucidate their value to educational practice.

- The purpose of a third proposed study, a research community culture analysis, is to communicate the richness and productivity of the community of researchers that has evolved, at least partly, because of funding that it has received from NSF. A significant number of people have served on NSF-funded projects and have gained knowledge and experience while working on those projects. The work and expertise of these researchers and others extend beyond the boundaries of the work that they have performed for the NSF. An analysis of this community can reveal some of the extended effects of NSF programs.

- The fourth proposed study, generalizability analysis, is an attempt to attend to the spectrum of impacts that NSF programs can have. The analysis would use sampling techniques and large-scale instruments to produce information about results from a collection of funded projects, and the analysis would attempt to identify the impact of those studies upon likely users.

The body of this paper begins with a brief description of one NSF program, Research in Teaching and Learning (RTL), to exemplify the complexity of a funding program and the wide variety of projects that are funded. The other programs that are pertinent to this study, such as the Applications of Advanced Technologies Program and the Educational Indicators and Studies Program, have comparable characteristics and are equally diverse. The description of the RTL program is followed by a discussion of the diversity of research in education. This discussion is followed by statements of specific evaluation questions that are central to this kind of undertaking, and by a brief enumeration of issues and pitfalls that are likely to arise in the evaluation of NSF programs. The paper concludes with an outline of four promising approaches to program evaluation that would communicate the value of NSF programs to the most important audiences that could use the results of this kind of evaluation.

Brief Description of the Research in Teaching and Learning Program

Overcoming conceptual difficulties in science; generating more and better mathematical discourse in elementary classrooms; building models of student achievement in science and mathematics; identifying the theoretical and national policy implications of the persistence of high-ability minority youth in college mathematics, science, engineering, premedicine, and predentistry programs; and assessing changes in home processes related to children's interest and proficiency in mathematics as they are affected by a program designed to help parents to be more active in their children's mathematics learning: these are only a few examples of the 187 new and continuing projects that were supported by the Research in Teaching and Learning Program during the 1987-91 period. Over these 5 years, grants totaled $23.45 million—not including the funding of 26 projects that were shared jointly with other NSF programs between 1987 and 1990. Significantly, in terms of the numbers of projects funded, the greatest concentration of awards was in the field of mathematics, followed by physics, general science, interdisciplinary area, biology, chemistry, and astronomy.

Goals of the Program

Research in Teaching and Learning—a program in the Division of Research, Evaluation, and Dissemination—seeks to support new discoveries about how individuals and groups learn, teach, and work more effectively in complex, changing environments. RTL supports basic and applied research to answer questions about the teaching

> *"Research in Teaching and Learning... seeks to support new discoveries about how individuals and groups learn, teach, and work more effectively in complex, changing environments."*

61

and learning of mathematics, science, and technology at all levels. Findings from this research are to inform those who are active and interested in education and its reform. Policy makers, teachers, teacher educators, curriculum developers, parents, and researchers are among the people who compose the intended audience for the research output and findings that appear in reports, videos, computer software, laboratory activities, and instructional materials. Although RTL has been supporting research since 1984, its current priorities are to advance our understanding of the following:

- How students learn complex concepts in science and mathematics;
- How advances in knowledge of mathematical modeling link to the learning of complex concepts in science;
- How teachers' subject-matter knowledge and competencies affect student learning; and
- How teachers learn to become inquiring practitioners and active researchers, and how they learn to apply that knowledge in their classrooms.

The goal of the RTL program is to generate a knowledge base that informs the national movement to reform mathematics and science education. To attain this goal, the program has specific objectives.

- First, the program seeks to establish the content and sequence of learning that can be most effective in developing science and mathematics literacy and problem-solving skills.
- Second, the program endeavors to meet the current and future needs of decision makers and other people who perform critical roles in education and research by building a coherent and comprehensive base of knowledge of learning and teaching in mathematics, science, and technology.
- Third, RTL seeks to produce research that will inform the reconceptualization of performance measures and that will develop alternative methods for assessing student learning.
- Fourth, the program is to study the significance of the nature and quality of laboratory experiences and determine their effects.
- Fifth, RTL is to explore factors—especially those influencing underrepresented groups—that empower students to participate and achieve in science and mathematics and to develop a positive disposition toward these fields of study and work.
- Sixth, the program seeks to engage teachers in education research, as a strategy to help make findings become more closely attuned to classroom reality.
- Finally, RTL's seventh objective is to assure that research findings are applied by members of the education community—teachers, teacher educators, policy makers, educational administrators, parents, and other researchers.

Range of Projects Funded by the Program

Projects supported by the Research in Teaching and Learning Program vary in their purposes, methods, age levels of student populations, and subject matter. As indicated by the nature of the projects that were cited at the beginning of this paper, goals of projects can range from addressing policy issues and providing information for policy decisions to very specific learning problems. The program uses five categories to group and describe the range of its projects: setting the research agenda, research in teacher enhancement, research on student learning, curriculum research, and cross-cultural research.

RTL involvement in setting the research agenda includes supporting major conferences, reports, and publications within the research community. Recent funding has been directed toward research projects that advance current efforts to reform mathematics and science education. For example,

- The "NCTM Research Catalyst Conferences" had six groups of researchers, each of which involved two mentor researchers who met with less experienced researchers, to design and encourage research critical to the implementation of the National Council of Teachers of Mathematics (NCTM) Curriculum and Evaluation Standards for School Mathematics.
- Another RTL-funded project prepared the aptly titled report "Establishing the Research Agenda: The Critical Issues of Science Curriculum Reform." This report was discussed at national meetings and published in the *Journal for Research in Science Teaching.*

Other funded projects have helped to define the research agenda in education by summarizing key research findings and by examining ways that findings can be communicated to practitioners.

Funded research in teacher enhancement targets the teaching process and reveals ways that student learning of mathematics and science can be expanded.

- The Cognitively Guided Instruction Project, directed by Elizabeth Fennema and Thomas Carpenter at the University of Wisconsin-Madison, and funded jointly with the Division of Teacher Preparation and Enhancement, produced research-based materials and strategies for inservice and preservice teachers to be more effective by using knowledge about student thinking to make instructional decisions.
- Another example of funded research in teacher enhancement is a school-based research project that is run cooperatively by the University of Maryland and the Montgomery County Public Schools in Maryland. Project Impact (Increasing the Mathematical Power of All Children and Teachers) strives to enhance student understanding of mathematics through summer inservice programs for teachers of minority children.

In the course of these programs, teachers study pedagogical content knowledge, mathematical content knowledge, and their beliefs. Teachers use the opportunity to examine and develop

"... goals of projects can range from addressing policy issues and providing information for policy decisions to very specific learning problems."

63

instructional activities that foster mathematical understanding and problem solving. Evaluation is ongoing in studying the implementation of the summer inservice goals and in a multiyear impact evaluation of the effects of the inservice programs on student learning and teacher beliefs and practices.

Research on student learning embraces projects that focus on student cognition, concept learning, problem solving, and the knowledge that students bring to the formal educational setting.

- In science, funded projects are devising and studying new ways to help students learn such traditionally difficult concepts as force, motion, gravity, harmonic motion, and the adaptation and natural selection mechanisms that underlie biological evolution.

- In mathematics, funded projects focus on topics that range from early number concepts through multiplication, estimation, prealgebra and algebra, geometry, calculus, probability and statistics, and abstract algebra at the college level.

Curriculum research includes projects that endeavor to inform instructional materials development. Research focuses on topics from the school and college curriculum and is designed to foster curricular and instructional innovations.

- In science, research on topics in physics, chemistry, and biology help to structure instructional materials.

- In mathematics, other studies focus on Logo geometry for elementary schools, mathematical modeling and exponential functions for high schools, and calculus concepts and computers for courses at the college level.

Finally, in funding cross-cultural research, RTL intends to raise the expectations of educators concerning student achievement and classroom practices by studying practices and results from other countries.

- The work of Harold Stevenson and others on Japanese, Chinese, and American students has been highly acclaimed and widely published in both the scientific and popular press. These researchers have articulated their objectives in the following terms, "the goal of this research is to increase understanding of prior and contemporary influences on achievement in mathematics so that effective suggestions may be made for the improvement of mathematics education in the United States" (*NSF Summary of Awards, Research in Teaching and Learning, Fiscal Years 1987-1990* [hereinafter NSF 1987-90] 80).

- Other research projects in this category seek to maintain and enhance the database of the IEA Second International Mathematics Study and provide American educators access to research monographs that have been published exclusively in the former Soviet Union.

The five categories of research projects that are supported by the Research in Teaching and Learning program embrace diverse projects. The objectives of the projects that are included in these categories range from very broad issues of reform and international perspectives to very specific concerns in concept learning, teaching practice, and materials development. Moreover, the range of project goals within any given RTL funding category is very extensive and broadens, rather than concentrates, the diversity of research endeavors. For example, under the category of research on student learning, some studies use students' mathematical errors as a springboard to critical thinking (NSF 1987-90, 6); another project focuses on systems of concepts in multiplicative structures; a third studies the cognitive processes that are involved in understanding and using scientific diagrams; and still another project is attempting to facilitate the process by which students learn to connect real-world phenomena with scientific representations of the phenomena. This variety in projects is also evident in the educational level that serves as the focus of funded research, as indicated below.

- Approximately 29 percent of the projects during 1987-91 concerned the elementary level of education;

- Fifteen percent concerned the middle school level;

- Thirty-three percent concerned the secondary level;

- Eighteen percent concerned the undergraduate level, and

- Nineteen percent were not related to any single grade level, since some projects treated more than one grade level.

Another indicator of the diversity of these projects is the fact that over 300 key words and phrases are listed in the index of the 1987-90 RTL summary report. In brief, NSF's program of Research in Teaching and Learning appears to seek broad coverage over concentration in the projects that it supports, since RTL addresses learning and teaching by people of all ages, and since RTL tries to provide information for decision making by a range of people, including parents, teachers, administrators, scientists, policy makers, and curriculum developers.

The Practical Nature of Research in Education

Educational research incorporates many kinds of inquiry and is not limited to a particular mode of investigation. The objectives of educational research can range from efforts to understand the learning process to the gathering of information that is intended to improve decision making. Borg and Gall (1983) have classified educational research into four typologies that differ according to the following characteristics of the research: topic, purpose, hypothesis testing, and basic versus applied research.

Topic describes the phenomena investigated, such as learning process, cognitive abilities, and teaching methods. Purpose addresses whether the research attempts to describe or characterize a group of phenomena, or whether it tries to reveal relationships among variables. Hypothesis testing research involves studies based on some prior theory or findings that are used to confirm or reject conjectures. Basic versus applied research distinguishes between research that focuses on understanding fundamental structures and processes (basic research) and research that focuses on structures and processes as they appear in educational practice (applied research).

Research to Inform the Practice of Teaching and Learning

Although the nature of educational research is varied, education is a practical field that continually requires teachers, administrators, supervisors, and others to make decisions that have cumulative influences on the lives of students. Research that facilitates decision making, that provides guidelines to help reduce the complexity of educational content and instructional practices and materials, or that provides answers to questions that arise repeatedly has enormous potential for teachers and others, provided findings are put in a useable form. For example, knowing that 6-year-old students enter first grade with thinking strategies that are useful when solving mathematics word problems that have generally been presented to older students (Carpenter and Moser 1983) is a powerful finding that could help first grade teachers to work effectively with their students.

The curriculum, the goals, objectives, and instructional materials that are necessary to achieve desired outcomes, is a dominant force in determining what is taught in classrooms in this Nation. Research can benefit the development of curriculum materials by indicating what works and what does not work. Systematic feedback on draft versions of instructional and curricular materials can be critically important to curriculum developers who are writing materials for use in classrooms.

"Research can benefit the development of curriculum materials by indicating what works and what does not work."

Research to Lead Reform

The relationship of research to education reform often incorporates an important bifurcation: research can prompt reform; or research can be a response to reform. To cite a significant example, the NCTM Standards were written by people from the research community and by other mathematics educators in the profession who were very knowledgeable about research findings. This knowledge—in addition to collective, accumulated experience in teaching and producing effective curriculum materials—was very valuable in the preparation of the NCTM Standards that have served as a driving force in current efforts to reform mathematics education in this country. Furthermore, the NCTM Standards went beyond existing, verified knowledge and established new expectations regarding the nature and extent of mathematics that all K-12 students should experience. The Standards also presented content topics (e.g., discrete mathematics) for which there were very few available curriculum materials. The Standards then set an agenda for additional research that would be needed to effect the vision that the Standards offered to the community of mathematics educators and researchers. In this manner, research can do more than add fuel to the fervor for reform by helping to ignite the flame and by adding tinder that will keep the flame going.

Research to Develop and Confirm Theories

Theory building, theory verification, and model building have been applied to education and have been an application of research. If we share the view of Kaplan that "a theory is a way of making sense of a disturbing situation so as to allow us most effectively to bring to bear our repertoire of habits" (1964), and that a model is "the embodiment of a structural analogy" (1964), then we can see that theories and models are useful in providing a language for communication and in making predictions. Indeed, with well-developed theories and models, predictions can be very precise. Piaget's theory of the development of intellectual capacity in children, and its focus on their attempts to structure their world and give it meaning, fostered a large body of research. Carroll's model of learning (1963) that depicted learning as a function of prior knowledge, perseverance, opportunity to learn, and other variables, was instrumental in the mastery learning movement and was used to design research to verify that model under different conditions. Because education is complex and involves many variables, educational theories and models are difficult to develop, but successive iterations in the development of these theories and models help to define research questions more precisely and productively, and link individual research studies to other bodies of organized inquiry.

Research to Explain Outcomes and Practice

A common use of research in education is to describe outcomes, practices, and conditions. Teachers who are isolated in their classrooms can benefit from descriptive studies that reflect on the practices of others. Such studies provide confirmation for a teacher who rarely has the opportunity to observe other teachers in their classrooms or to consider variations in teaching practices. National and international studies that describe the achievement level of large groups of students, or the achievement differentials by different groups of students, are helpful for policy makers when they review policies and allocate resources.

Education is notorious for borrowing direction and methods from many other fields, such as psychology, the natural sciences, anthropology, and linguistics. Educational research is no different and has applied a variety of methods to study questions that bear on the field. The range of methods includes ethnography (anthropology), computer simulations and models (computer science), case studies (medicine and sociology), statistical analyses (statistics), cost-benefit analyses (economics), and policy and historical analyses (political science and history). These methods of inquiry have an impact on the ways that researchers interact with their findings, and can reveal different information concerning the same phenomena. In light of the large number of variables, factors, and complexities that arise in most educational research, multiple methods of research are necessary if we are to begin to identify and understand the important variables and relationships among variables that exist in education.

Research in Science and Mathematics Education

The nature of research on teaching and learning in science education and in mathematics education is defined by a multitude of factors. In a certain sense these fields are very young. The bodies of knowledge, methodologies, and traditions that they draw upon are continually under development. Moreover, both fields are greatly influenced by the content areas of mathematics and science, and many researchers have been trained in those disciplines. In addition, the education of students in these content areas requires attention to psychology, learning theories, and educational foundations. Because education is so diverse, researchers in both science education and mathematics education have drawn upon many methodologies to study teaching, learning, curriculum development, and policies. The emerging technologies and their applications to education have required mathematics and science education researchers to expand their knowledge to understand these new and changing forms that have the potential to change drastically the teaching and learning of mathematics and science. Given these varied sources and methods in education, research on teaching and learning in mathematics and science education calls for corresponding variety in the approaches that are used to conduct research and maintain contact with research advances. This compounding of complex educational methods and research approaches often makes it difficult to understand the research, and to identify and communicate the value of the research.

"The communication of the value of the programs requires depicting what the programs have done, what their main effects are, and how these effects have been applied to practice."

Evaluation Questions

The questions that are to be answered in the course of evaluating such NSF programs as the RTL program should be structured by the purpose of the evaluation. As argued at the beginning of this paper, the central purpose of NSF program evaluation is the communication of the value of NSF programs that fund research and innovative technical applications in mathematics and science education. Communication is constructing knowledge. The acts of writing, speaking, reading, and listening require building on existing knowledge, making decisions, analyzing information, and drawing conclusions. The act of communicating the value of NSF also entails constructing the value of the programs by focusing on what is important, analyzing information, and drawing conclusions. The communication of the value of the programs requires depicting what the programs have done, what their main effects are, and how these effects have been applied to practice. But the communication process also attends to an audience and sends a message. As a central purpose for evaluation, the communication of the value of programs combines the substance of the message with the message itself.

Clearly, additional purposes for an evaluation of the effects of an NSF program can be phrased in other ways. One purpose could be to ascertain the accomplishments of the RTL program and the impact of these accomplishments on instruction and learning in mathematics and science in the United States. Two other purposes could be served by the evaluation: the undertaking can gather information targeted to strengthen the program, so that it will be more effective in achieving its goals; and the evaluation can reflect upon the goals of the RTL program. In reflecting on the goals of the program, attention would need to be given to their relationships with the goals of other programs, so that a clear view of the correspondence among goals can be obtained, and so that the future needs of mathematics and science education over the next 5, 10, and 15 years can be defined. In brief, the evaluation of the program needs to be specific enough that it can be accomplished, but it needs to be general enough that it will provide confirmation, direction, and a rethinking of procedures. Focusing on the communication of the value of the program can meet these criteria.

Sample Questions

In communicating the value of NSF programs, there are at least six important questions that the evaluation should strive to answer.

1. What research findings and information have been produced by individual projects and by the collectivity of projects that have been supported by the RTL program?

In thinking about this issue, it is useful to decompose the question into its components by employing the two-by-

Exhibit 1

Four questions for an evaluation of the Research in Teaching and Learning program, structured by the information that is now known as a result of the research and by the applications of those research findings.

Research Results	Applications	
	Yes	**No**
Know	What findings and information have been produced that have successfully solved a problem or fulfilled a need?	What findings and information have been produced that have not been applied to solve an important problem or fulfill a need?
Do Not Know	What critical problems or needs have not been resolved or refined by research findings and information?	What negative or poor applications have filled the gap in the absence of solid research findings and information?

two matrix that is depicted in Exhibit 1. One dimension represents the information and findings that have been produced by RTL projects. This "research" dimension can be divided into two categories—what we know and what we do not know. The second dimension represents the application of research to existing problems. This "application" dimension can also be divided into two categories—what research has been applied and what has not been applied. This simple matrix helps to generate four classifications of questions that should be answered by the program evaluation.

1a. What findings and information have been produced that have successfully solved a problem or fulfilled a need?

The responses to this question will be the success stories of the program. Projects that have been successful in gaining results and in having these results applied to the solution of important problems will provide strong evidence about the impact of the program. An important part of the answer to this question lies in the identification of problems and needs and in demonstrations of the ways that funded research provided solutions to the problems or met the needs. In addition, it is critically important that the question and consequent answers focus on significant problems. For example, helping elementary teachers to learn how to build on student thinking in their teaching is more significant than deciding between the use of vertical addition or horizontal addition.

"*The evaluation should determine what the program has not done in areas where information and research results would be useful.*"

Clearly, assigning importance to problems is a value judgment, and that reality should be considered in the design of any evaluation.

1b. What critical problems or needs have not been resolved or refined by research findings and information?

The evaluation should determine what the program has not done in areas where information and research results would be useful. Some explanation of why research has not been successful in resolving—or at least, in refining—important problems will need to be incorporated into the answers to this question. There may be many important reasons why research findings are not available. Possible explanations might include the following: research may have been tried, but findings may have been inconsistent; funds may not have been available to support the needed research; the research may not have been concentrated in the manner that would have been most likely to resolve the problem; and insufficient time or resources may have been allocated to solve the problem.

1c. What findings and information have been produced that have not been applied to solve an important problem or fulfill a need?

In any research program, some efforts will not produce the intended results or will not be productive. Alternately, some research may not address questions that are as important as other research. One would hope that there would be a minimum of such research that is supported by the RTL program. However, a program without any such efforts is probably insufficiently aggressive in advancing knowledge in a given area. Still other research will address basic questions whose answers do not have any immediate applications. For example, some psychological research in the learning of nonsense syllables is basic and lacks direct classroom applications. A complete evaluation of the RTL program would need to identify research efforts and findings that have not been applied and would need to assign some value to these efforts, since they may have made a significant contribution to a body of knowledge and may be an important outcome of the program.

1d. What negative or poor applications have filled the gap in the absence of solid research findings and information?

Any program that supports research will have to decide between the research that it will fund and the research that it will not fund. In some instances, important educational questions will arise, and no information from research may be available to help respond to those questions. The absence of this information may suggest that the program has failed to anticipate the issues that will arise in the future. In that event, practitioners will have to use the best information that is available to them. In some cases, the information that is available or the practices that are current may be relatively unsuccessful or may even produce poor results because the needed information has not been produced. For example, some feel that an extended use of mathematics worksheets with young children can result in rote learning and the development of a very mechanistic view of mathematics. Without research findings that refute this practice, some teachers will continue to have a worksheet-based mathematics classroom. Consequently, the evaluation of the RTL program should at least acknowledge the kinds of research that have not been funded and should consider the implications—both positive and negative—of the decisions not to fund certain research.

In addition to evaluation questions that focus on the application of research findings, there are five other questions that should be considered.

2. How has the RTL program contributed to the development of a community of researchers who serve as resources for the education system?
3. How have findings and information from the RTL program supported other program efforts, and how have the findings and information been used by other NSF programs, such as that in Instructional Materials Development?
4. How has the RTL program shaped and set the research agenda in mathematics and science education; and, more particularly, how has this agenda setting derived from provocative questions that have been formulated by the program and that have motivated large numbers of studies?
5. How have the RTL program and its funded projects built on findings from related research programs and fields, such as those in psychology and computer science, to ensure that supported research is relevant and does not duplicate work in other fields?
6. How have the operations and funding strategies of the RTL program served the program's goals?

Issues and Pitfalls in Evaluating the Research in Teaching and Learning Program

There are seven issues that are central to the design of program evaluations for the National Science Foundation.

- One issue concerns the unit of analysis for an evaluation. To show fully the extensiveness of the NSF program's accomplishments, whenever possible, the unit of analysis should be the program.
- Scale is a second issue. One major goal of the NSF is to improve the quality of the Nation's science, mathematics, engineering, and technology education. Trying to observe movement in the national system poses massive problems for the comprehensive evaluation of programs.
- A third issue is that the observation of important effects will depend somewhat on the time and duration of the research projects. Sometimes important systematic effects do not appear until years after the completion of a project. Also, the research or project could have been worthy, but the project or research may not have been extended over an adequate period of time to produce observable systematic effects.
- A fourth issue is that change and the evidence of change is not uniformly apparent over the education system. The problem becomes one of locating the points at which change has been concentrated in the educational system, and of attributing the change to identifiable research and development projects.

"Sometimes important systematic effects do not appear until years after the completion of a project."

- A fifth issue concerns the synergy of the research and education systems and how information flows between the two. Funded research may be of a high quality, but the dissemination of findings may be poorly implemented.

- A sixth issue in studying the impact of research on practice is that there may be conflicting forces that bear on the support of research and the application of research. What research has determined to be theoretically sound practice may confront current practice that is strongly embedded in tradition and values. Or, the recommended changes may be overwhelmingly expensive. Quality research cannot always be expected to find its way into practice.

- Finally, in any evaluation of research programs there are unintended outcomes that in many cases will be positive. This implies that in studying NSF programs some consideration needs to be given to effects that go beyond those stated in projects' proposals or final reports.

> *"... in studying NSF programs some consideration needs to be given to effects that go beyond those stated in projects' proposals or final reports."*

Promising Approaches to Evaluation

Evaluating the impact of the NSF programs is complicated, as indicated earlier, by the great variety of projects that were funded under the programs, the range of age groups that were targeted by projects, the forces within education that retard the implementation of research findings, and the lack of concentration of results that can be brought to bear on the educational system in the United States. Tracking the effects of any one of the programs, such as the RTL program—on the profession, on other research, on practice, and on institutions—is further complicated by the many other influences that affect schools and education. Alternate approaches to evaluation are needed in order to reveal the levels of outcomes and the variety that exists among outcomes. In light of these considerations, some nontraditional approaches to evaluation can communicate to others the value of NSF programs that fund research and innovative technical applications for mathematics and science education. To help simplify references to the different programs, the four approaches to evaluation are described in the context of only one funding program, Research in Teaching and Learning. The approaches, however, could be applied to any of the other programs or to combination of programs.

Retrograde Analysis

One indicator of a research program's value is its internal integrity: how research produced over the years builds upon research that was previously produced by the program. A program with internal integrity will develop a coherent body of knowledge with evident chains of inquiry. The value of the program, in this case, is the created body of knowledge that can be drawn upon by different people for multiple reasons. Strong chains of inquiry are more apt to lead to significant applications when ideas are highly developed, expended effort and resources have been concentrated, and findings have stood the test of time. Communicating the value of created bodies of knowledge becomes a problem of describing what the body of knowledge is, how it has evolved from the work of projects within the program, its importance, and its potential applications.

The study of the internal integrity of the RTL program and the bodies of knowledge that it has generated could be done by a team of three people—one evaluator, one science educator, and one mathematics educator. The principle charge to the evaluation team would be to analyze the relationships that exist among the findings of projects that have been supported over time. The central focus of the evaluation would be to document the relationships among the findings of the most successful projects and to establish the fact that projects have built on each other to form coherent bodies of knowledge. The work of other projects could be studied as appropriate or warranted. The most productive projects to begin this investigation can be identified from the amount of funding received, the visibility of the project, and the extensiveness of findings. The Cognitively Guided Instruction (CGI) project is one example of such a "star" project.

Instead of the usual approach to evaluation, which examines the progression from early studies to more recent studies, it would be useful to proceed in a retrograde manner, by examining the ways that more recent studies have relied upon and built upon a succession of earlier studies. Such retrograde analysis would examine relationships between funded projects by focusing upon the "generation" of the projects under consideration —by moving from the current research generation to research that was funded and conducted one, and two, and three or more generations earlier. In this approach to program evaluation, what is currently known from each of the "star" projects could be described by using information obtained in interviews of the project staff and others, by reviewing project documents and technical papers on findings and results, and by surveying other sources of information as appropriate. Then, one could analyze the research bases for the current findings and information, and the derivation of these research bases from research that was conducted one and more generations earlier. In this manner, a project genealogy would be produced (Webb, Shoen, and Whitehurst, 1993). Subsequently, the linkages between research generations would be used to identify the initial or formative ideas that underlie research over time. The intent in this approach to analysis is to establish a chain of inquiry linking the generations of projects, and to relate this chain to support from the RTL program or to the manner in which RTL has built upon support from other sources. Such an analysis has the potential to demonstrate the cumulative or building effect of research findings, the evolution of projects over time, the evolution of project staff thinking, and the productive use of RTL funding. The most likely chains to be revealed are ones that follow a researcher, group of researchers, or a topic of research. Theoretical mappings and idea tracings over time are possible outcomes.

Chains of inquiry and other findings from this analysis can be validated by direct evidence—a researcher reporting and showing evidence of a link to work of another project—or triangulation of evidence—confirming evidence received from different sources. The final product of this evaluation could be a report including both a narrative explanation of the linkages found and charts depicting the development of bodies of knowledge.

"Instead of the usual approach to evaluation, ... it would be useful to proceed in a retrograde manner, by examining the ways that more recent studies have relied upon and built upon a succession of earlier studies."

Video Documentation

A second evaluation approach to communicating the value of the RTL program builds on Marshall McLuhan's idea that the medium is the message. The central evaluation question focuses on the coherent messages about classroom practices and educational innovations that can be gleaned from the program. The form of reporting findings from this investigation would be a video documentary. The process of creating the documentary will be, in and of itself, an evaluative investigation extended further by using the different elements available in video to communicate the findings. Video is a powerful medium for reporting to large and varied audiences. Video, as compared to text, has the advantage of communicating more clearly the visual materials that are produced by projects, new applications of technology, and the full range of diverse projects that form the program.

The preparation and production of the video RTL documentary would be the responsibility of an evaluation team consisting of an evaluator, mathematics and science educators, a producer, a script writer, and necessary production support staff. The time that any one person would spend on the evaluation would depend upon the extensiveness of the study and the role to be assumed. The process would begin by researching and analyzing the main messages that can be derived from the RTL program. Then, the selection and focusing process would identify the major theme or themes for the video, based on validated findings, what has been put into practice, what is visually exciting, and what is ongoing, exciting work that has the potential for change. Subsequently, an editorial board, consisting of NSF staff, researchers, and others, would critically analyze the themes and the work selected to create the video and to substantiate the selections of material. The evaluation team would need to have some autonomy to do the necessary research, prepare the script, and produce the video. Some written materials could be prepared in support of the video, but the video should be the main form of communication.

The actual story and the major themes of the documentary will be decided as part of the process of evaluation. Many possibilities exist.

- One is to report on actual classroom applications where practices have been directly influenced by RTL projects. A variation in focusing on classroom practices would depict the applications of research findings by making a composite of an ideal classroom for different grade ranges and content areas. Classroom composites could reveal in concrete terms the practical body of knowledge that has been generated by funded research. The classroom composites could consist of written and video scenarios of the RTL-influenced classrooms that depict teaching practices, student activities, and student learning.

- Another possibility for the story line of a documentary would be to take an issue, such as opportunity to learn, and show how RTL projects, such as the Second International Mathematics Study (SIMS), have advanced our understanding of that concept and how there are consequences that can be documented or anticipated from this advancement. For example, SIMS data indicated that opportunity to learn was strongly correlated with achievement, as has been supported by other studies. This can be a powerful message when thinking about "world class standards." A treatment of opportunity to learn could also lead to a timely analysis of equity, and to an analysis of differences in content and in presentation to various groups of students.

An evaluation of the value of RTL and other programs would grow out of the process of revealing the implications of what we know to be true and what we think is possibly true.

In addition to investigating major themes across the RTL programs and their applications to practice, the video development process can be used to reveal the questions that projects are pursuing and the substance of what is being learned. Many projects use video as a research tool to record student interviews and classroom interactions, and a video documentary could build on these video resources that communicate very well what has been developed. This could be accomplished by collecting video and other visual materials from projects, by abstracting depictions of new findings and applications, and by creating video episodes to present the major ideas. This process serves both as a means of evaluating the richness or weakness of findings and as a form of communicating and describing some of the RTL program's effects.

Other video techniques afford unique ways of communicating the range of findings, the scope of work, and the applications to practice. Some of these techniques are:

- Video interviews with researchers, teachers, and students;
- Voice-over segments that illustrate a new practice while the audience hears a teacher reflect on the practice;
- Montages that present a range of investigations through a sequence of music-accompanied images that are flashed on the screen; and
- Presentations of computer simulations, software demonstrations, or CD-ROM applications to explain the wide use of technology that is being supported by RTL.

The process of producing a video documentary using these and other techniques, along with presenting major themes and applications, requires grouping RTL work and findings into categories, deducing meaningful conclusions, portraying classroom applications, and validating what is reported. All of these activities are part of an evaluation process and communication.

"Formal review mechanisms can be imposed on the development of the video to ensure that the substance of reports and communications adheres to the rigorous requirements of good evaluation."

Formal review mechanisms can be imposed on the development of the video to ensure that the substance of reports and communications adheres to the rigorous requirements of good evaluation. A review process can be designed to include an editorial panel, researchers as advisors, and practitioners. These people would have the responsibility of ensuring that the information presented is accurate, and that the information appropriately communicates the effects of the RTL program and how the findings benefit the educational system. Outside reviewers can be employed as impartial technical advisors and even on-screen critics or discussants. Cost controls would need to be imposed, but the expense of developing a 30 to 60 minute video documentary of studio quality could be less than the cost of developing both a conventional study with similar evaluation purposes and a video that describes the study's findings. It is likely, however, that the cost of a video documentary will vary with the overall quality of its content and imagery. The least expensive video would derive from a collection of existing video materials from RTL projects; the video and evaluation would be edited from these materials and presented with a sound track to communicate the range and value of RTL projects. The most expensive video would consist of original footage; the video would be of network quality and the analysis would investigate RTL's impact on classroom practices.

Research Communities Culture Analysis

An important contribution of the RTL program and other NSF programs is the development of mathematics and science education research communities. A cultural analysis could be carried out on these communities and on the links that these communities have with other relevant professional communities. The analysis of the mathematics and science education research communities could then be compared or contrasted with analyses of research communities in other subject-matter areas (such as language arts, social studies, and fine arts), other funding situations (such as the private sector or research funded by private foundations), or in other countries.

An evaluation team would be responsible for conducting the analysis. This team would—at a minimum—be composed of a mathematics educator, a science educator, and a cultural anthropologist/evaluator. In exploring the culture of researchers that has evolved through their individual interactions with the RTL program and other NSF programs, a number of questions can be addressed.

- What constitutes the research community culture that has evolved through NSF programs? Which people form the community? What is the entry into this community and how do people drop out?

- What interactions exist among the members of the community, when one considers both the mode and frequency of interactions? How do members of the community join together for cooperative work?

- What beliefs are shared by the members of the community? What support systems are in place?

- What are the patterns of migration and grouping? What are the traditions and forms of communications? Is there a common language? Are there those who would be considered outliers in the community?

- What alliances have been formed with the community and other organizations and groups? What is the power base within the community, and how powerful is the community in relation to other communities?

- What is the "gross community product" as indicated by materials produced, conference presentations, funding generated, and other measures of production?

- What are the mechanisms for transmitting the culture, and is it in the process of expanding or declining?

The main sources of information for a cultural analysis would be the researchers who have received funding through NSF and others who could be considered members of the culture (graduate students and other researchers closely aligned with members of the research community). A cultural analysis would gather information from the members of the community using interviews, questionnaires, personal resumes, and other sources used by anthropologists in studying cultures. One of the fundamental questions that would be have to be addressed first in such a study concerns the actual existence of communities of researchers in science and mathematics education. Even though communities that are identified may not be considered to be "cultures" from a narrow anthropological perspective, such an analysis could produce useful descriptive information about the communities that will communicate some of the value that has been gained through the NSF programs. The methodology of cultural analysis, as used by anthropologists and others, offers the means to validate the information and conclusions that would be developed in such a study. Contrasting the research communities that have evolved out of NSF programs with other situations where other research communities have —or have not—evolved would add to the credibility of information about the importance of NSF. For example,

- One significant point of contrast might be found in the educational research communities that exist in other countries, a contrast that would be instructive despite acknowledged differences between educational systems and their relationships to local and national government.

- Another significant contrast might be found in the research communities that have formed in this country for other content areas in which no NSF funding is available.

- A third significant contrast might be found in the work and interactions of educational researchers who are supported primarily by private foundations, and in the interactions or overlaps of this group with the community of researchers funded by NSF.

- Yet another source of confirming information would be to consult the different mathematics and science education professional organizations, to ascertain the value placed by these organizations on the research communities at issue. Some indicators of this value include the visibility of the research communities in these organizations and the distribution of research findings by these organizations.

The ultimate product of the culture analysis recommended here would consist of written reports that would provide detailed profiles of research communities, their relationships with NSF, their contributions, and their uniqueness in contrast to other research communities.

Generalizability Analysis

In order to identify and examine the breadth of the RTL program's impact it would be useful to undertake a generalizability study. The purpose of a generalizability study would be to consider the impact of the program by looking at a sample of projects that have been selected randomly from those funded by the program. Although the study would reduce the costs of studying program effects by focusing on a smaller number of projects, it would have the power to suggest generalizations about the program. Certainly, the ideal situation would be to be able to study, in depth, all of the projects that have been funded by the program, and to report the effects of

---

"The purpose of a generalizability study would be to consider the impact of the program by looking at a sample of projects that have been selected randomly from those funded by the program."

"The four varieties of studies that have been described in this paper have been designed to provide information on a range of effects of NSF programs."

each one. However, with the nearly 200 projects funded by RTL, for example, this would be a very large and expensive task. One assumption for doing a generalizability study is that it is important to look at the effects of the program as a whole, rather than the effects of only a few projects that might be considered to have been the most productive. One reason for doing a generalizability study is that not all projects have the same scope or concentration as others. Some projects serve specific local needs; others support beginning researchers; and others may be in the very initial stages of developing an important chain of inquiry. A random sample of the projects from a program would provide a cross-section that would offer a better description of the whole program than a review of only a few, large "star" projects. How large a random sampling is needed and how the selection should be done would depend on the program and the different facets of the program to be considered.

The study of each project would require data gathering to document the effects of the projects on classroom practices, teachers, theory-building, and other applications. The expectation is that the findings from this cross-section of projects will be distributed across all of the four cells in Exhibit 1. Depending on the findings across the projects studied, statistical techniques can be used to generalize from findings common to a number of the sampled projects to all those in the program. Some supporting information on the extent of the effects of the NSF program can be obtained by using the more traditional means of administering questionnaires to a random sample of the members of targeted groups, such as the teachers' professional organizations (e.g., NCTM and NSTA), classroom teachers, scientists, and mathematicians. The purpose of these questions would be to determine what awareness members of these groups have of the NSF programs' findings, their knowledge of the findings, and the degree of implementation. This more traditional approach to evaluation is recommended in the expectation that it may determine, at some level, the range of people who are being reached by information generated by the programs. For example, a number of people are probably at least aware of some of the findings reported by Harold Stevenson from his study of Japanese, Chinese, and American students. Adherence to assumptions and conditions for doing the statistical analyses will be used to verify the findings and conclusions. The results of the generalizability analysis would be presented in a written report.

Discussion

The four varieties of studies that have been described in this paper have been designed to provide information on a range of effects of NSF programs. The four studies have been conceptualized in nontraditional ways so that they could capture aspects of the NSF programs that may be overlooked by more conventional analysis and so that they can communicate the value of the NSF programs.

- The retrograde analysis can be used to examine the effects and value of research that emerges from within a program, and to communicate a clear view of how projects within a program build on each other. If the projects within a program do not build on each other, then it is very difficult to argue that people outside of the program will be using the results. Because the research efforts of a program are directed toward developing a body of knowledge, in the absence of some internal consistency the developing body of knowledge will be fragmented.

- The video documentary approach to evaluation can very effectively communicate to a wide audience the major themes and main messages that grow out of a program. The production of a video will depend on the existence of a created body of knowledge, but it will also consider applications of work beyond the projects that fall within a program. The process of producing a coherent and precise video requires a thorough analysis of the program under investigation. Video can be a very efficient way of condensing a large amount of information in a short period of time—information that communicates the range of projects that are supported by an NSF program.

- The cultural analysis of research communities focuses on the ways that NSF programs are developing a national resource of mathematics and science education researchers. A careful explication of these communities and the operations of these communities will document and probe one of the important contributions that the National Science Foundation has made. An analysis of clearly described research communities will highlight the work of these communities in producing research and applications under NSF sponsorship; simultaneously, the analysis will report the secondary effects of experience that has been gained through work on NSF projects, and it will identify the importance of those effects to other efforts—in teacher education, writing curriculum and evaluation standards, curriculum development, assessment development, and evaluation studies.

- The generalizability analysis is designed to reveal the spectrum of effects across an NSF program by studying a sample of funded projects. This kind of study can produce information on the range of research and innovation across a program, the diverse nature of these projects, and how these projects as a collection are infiltrating the educational system both locally and nationally.

Together, the four types of evaluation study treated here would present a strong profile of the National Science Foundation to its varied audiences, and would very effectively communicate the value of the Foundation's support of research and innovation.

References

Borg, W. R., and Gall, M.D. 1983. *Educational research: An introduction.* New York: Longman.

Carpenter, T. P., and Moser, J. M. 1983. The acquisition of addition and subtraction concepts. In *The acquisition of mathematical concepts and processes*, R. Lesh and M. Landau, 7-14. New York: Academic Press.

Carroll, J. B. 1963. A model of school learning. *Teachers College Record* 64: 723- 33.

Kaplan, A. 1964. *The conduct of inquiry: Methodology for behavioral science.* Scranton, Pennsylvania: Chandler Publishing Company.

Webb, N. L., Schoen, H., and Whitehurst, S. D. 1993. *Dissemination of nine precollege mathematics instructional materials projects funded by the National Science Foundation,* 1981-91. A final report to the National Science Foundation (grant MDR 9252727). Madison, Wisconsin: Wisconsin Center for Education Research, University of Wisconsin .

Footprints On Surfaces: A Nontraditional Approach To Evaluation Of National Science Foundation Programs

M. Christine Dwyer, RMC
Research Corporation

Introduction

This paper explores an approach for nontraditional evaluation of National Science Foundation (NSF) programs that deals directly with the impact of those programs on selected organizations engaged in education reform. The proposed approach advocates examination of a "slice" of the larger picture of educational change, focusing on selected stages and actors along the continuum from knowledge development to dissemination through implementation and reform. The examination would yield information about the stage linking the knowledge generated by NSF programs to implementation. The process would trace the influences and uses of that knowledge by intermediary organizations that have training and technical assistance functions, such as teacher training institutions, educational laboratories, and state departments of education. The basic idea of tracing influences on intermediary organizations is carried through in evaluation questions, variables, criteria for selecting a sample, and data collection processes. The paper illustrates the viability of the plan through an extended example and suggests some ways to address methodological problems.

This evaluation idea fits best with the purposes of those NSF programs designed to generate knowledge about the teaching and learning of mathematics and science to inform the work of researchers, policymakers, developers, and teachers. Several characteristics of NSF programs have inspired the design, including the following:

- The goals of creating a base of knowledge applicable to learners at all levels and useful to education reformers;
- The value placed on direct utility of projects for education;
- The targeting of underrepresented groups;
- The concern for systemic change;
- The variety of projects funded and the resulting array of outcomes and types of knowledge generated;
- The high profile among practitioners of many projects and their personnel;
- The collaborative nature of funded projects, which suggests multiple paths of project influence, and
- The emphasis on innovations.

Those characteristics also suggest the major challenges for evaluation design: the difficulty of capturing important, systemwide influences; the need for a new set of assumptions to replace traditional attribution concepts; the elusiveness of effects; and the need to separate development and dissemination for evaluation purposes. A study of the effectiveness of dissemination is not intended here. Lessons from the study of policy and program implementation over the past 20 years, along with our own

> *"The proposed approach advocates examination of a 'slice' of the larger picture ... focusing on selected stages and actors along the continuum from knowledge development to dissemination ..."*

experiences, have taught us that basing this evaluation on the programs' direct impact on educational practice would not be fair. So while this paper looks for connections to education practice, it is not intended and should not be interpreted as an evaluation of the dissemination or implementation of NSF projects.

"The purpose of the evaluation is to learn more about the varied paths and processes by which NSF programs influence educational practice, through a look at the impact on particular intermediary organizations ..."

In the next section, the evaluation purpose is discussed, with the goals of being fair to original NSF program intentions and also useful to policymakers. The section also includes an overview of the approach with special attention to explaining the concept of intermediaries. Following that is a summary of background influences that shaped the approach: the logical extensions of the Footprint metaphor; some applicable lessons from research about the influence of knowledge on policy and practice; and the author's experiences with the operations of technical assistance intermediaries. A framework for an evaluation plan, along with sample evaluation questions and a discussion of the nature of study results, follows. Finally, an extended example is presented, and practical issues to be encountered in carrying out the evaluation are discussed.

Key Features: Purpose and Rationale, Overview, and Role of Intermediaries

The purpose of the evaluation is to learn more about the varied paths and processes by which NSF programs influence educational practice, through a look at the impact on particular intermediary organizations that have the mission of linking research and practice for reform. The evaluation examines how the knowledge generated by NSF programs has affected or been incorporated by selected intermediaries within the larger education system. It focuses on those intermediary organizations with missions connected to systemic reform of mathematics and science teaching and learning. Simply stated, if knowledge was originally generated for the purpose of such reform, the question is how and to what extent active reformers have acquired and used the knowledge.

The proposed evaluation emerges from a "macro"-level perspective of how knowledge[1] changes practice, yet focuses on one element of the system of influences surrounding the knowledge generated by NSF programs. Instead of looking directly at effects on practice at the classroom or institutional level, it examines the effects on the larger system that supports, influences, and changes the work of education practitioners.

Intermediaries are agencies such as technical assistance centers, universities, teacher institutes, and laboratories with established dissemination, training, and reform functions. They serve both linking and leadership roles and bridge the cultures of research and development and educational practice through materials development, training, and networking. They are proactive in seeking knowledge generated by the research and development community. Intermediaries include the educational laboratories, the content-related Office of Educational Research and Improvement (OERI) research centers, technical assistance centers with categorical reform missions such as the 16 Chapter 1 Technical Assistance Centers (TACs), state departments of education, Federally supported project dissemination networks such as the National Diffusion Network (NDN), selected Statewide Systemic Initiatives (SSI), state and university projects for teacher training supported by the Eisenhower Mathematics and Science Education Program, universities that prepare teachers, and professional associations. Of greatest interest for this paper are those organizations with the closest connections to the reform of mathematics and science education.

---

[1] The term "knowledge" used throughout the paper is shorthand for the object of the evaluation—the myriad outcomes of project work, the ideas, principles, strategies, concepts, papers, curriculum manuals, software, materials, research results, etc., that form the work of NSF programs.

Evaluation Overview- The proposed evaluation would (a) illuminate the paths and processes by which knowledge generated by NSF programs is selected, acquired by, and transferred to intermediaries; (b) describe the knowledge that is of interest and not of interest to intermediaries; and (c) learn what functions that knowledge has served for intermediaries. Other possible evaluation purposes deal with the processes used by intermediaries to translate and transform knowledge and then the experiences of intermediaries in influencing education practitioners. The sample evaluation questions below suggest what could be learned from brief case histories of both intermediaries and the paths of influence of particular knowledge examples:

- How have regional Chapter 1 TACs used NSF-supported work in the teaching of elementary mathematics to improve programs serving disadvantaged students? Do the materials used by TACs include the principles and practices that emerged from the work on cognitively guided instruction, for example?

- To what extent do any techniques developed by specific NSF programs appear in the programs promoted and funded by the Department of Education's National Diffusion Network?

- Has Eisenhower-supported state-level teacher inservice been shaped by the knowledge generated by NSF programs?

The questions suggest the components of a model framework (i.e., objects, respondents, data collection processes) to bound data collection. Clearly, the evaluation process requires heavy involvement of at least some grantees and NSF in defining the information to be tracked, hypothesizing the varied influences of particular work on practice, and identifying the intermediaries that would be both likely and unlikely candidates for influence. Therefore, a component of the approach includes work with grantees to identify the presumed paths of influence of their work. The cluster evaluation method for identifying common outcomes would be relevant (Barley and Jenness, 1993) for identifying common paths of influence. The proposed data collection processes are akin to investigative journalism approaches (Smith, 1981; Guba, 1981), tracing leads about whether people in intermediate agencies are familiar or unfamiliar with, have used or not used, knowledge generated by NSF programs.

It is easy to anticipate arguments about this approach. One could argue that because the explicit intentions of NSF grantmaking did not (and should not) include the expectation of leaving traceable marks on practice, it is simply not valid to look for effects later. Or, from an instrumentalist perspective, one might assume that, because the intentions of knowledge developers may not have been specific uses of knowledge, it will simply be impossible to trace the processes by which that knowledge was acquired and transferred within the larger system. Finally, the anticipated elusiveness of information as a result of interpretation and translation over time may make the approach seem overwhelmingly complex to some.

On the other hand, it is very easy to imagine that policy makers and decision makers at all levels might expect an evaluation to answer the question of what and how and how much NSF programs have contributed to improved educational practices. The current climate of educational reform spurs everyone's interest in the extent to which changes in practice have occurred. The clear and widely promoted statements of needs for reform at all educational levels in mathematics and science teaching/learning and expansion of use of technologies have created a context in which there will be increasing pressures to look diligently for the mark of NSF programs directly on educational practice—and beyond, at student and societal outcomes. Further, because of the scope and depth of the current concerns about reform, one can anticipate pressures to look for those footprints on the "biggest surface" possible, perhaps even a national landscape—hence, the interest here in considering the larger systems that support the influence of knowledge on practice.

"There are several compelling arguments for using technical assistance and reform intermediary agencies as the 'surface' on which to look for footprints of influences from NSF programs."

More about Intermediaries- As is clear by now, intermediaries are a critical element of the evaluation design. Obviously, the value of using the approach depends partly on how possible it is to achieve agreement about which intermediaries act as primary channels or paths linking research and educational practice. Will grantees and NSF agree that it is both fair and valuable to trace and describe effects on the functions, understanding, beliefs, and attitudes of intermediaries? How complex will it be to attain agreement on which intermediaries are appropriate? While responses to those challenges are unknown to us at this point, it is a relatively simple matter to gather initial reactions. There are several compelling arguments for using technical assistance and reform intermediary agencies as the "surface" on which to look for footprints of influences from NSF-sponsored programs.

- First, from some perspectives, intermediaries represent manipulatable levers of change; in the spirit of systemic reform, it is critical to know how and to what degree they are influenced by and take advantage of the knowledge generated by NSF programs. They are likely to recognize and discuss the influences on their work, if any, of selected NSF programs because their espoused missions are to influence practice (whether by training, consulting, or product development) and to do so, they must be proactive in seeking knowledge and research.

- Second, intermediate agencies offer a potential solution to the problem of tracing isolated and discrete effects on practice and/or entirely avoiding looking at the effects on practice because of the complexity of where to look. Because of their multiple functions, intermediate agencies are likely to have had varied opportunities for contact with the knowledge generated by several NSF projects. For example, a regional educational laboratory initiative may have incorporated specific examples of technology use, as well as assessment practices and curriculum examples in its work with teachers.

- Third, because intermediaries are in the business of transforming research into materials, training, experiences, policies, and expertise for the purpose of influencing educational practice, they will be able to offer a rich perspective on the process of acquiring and using knowledge, and describing their own paths of contact and development, including how they have come to know and value NSF program material. Well-selected intermediaries would be expert reporters on the entire system of influences that connects knowledge generated by NSF programs to educational improvements.

- Fourth, depending on the intermediary, there may even be some limited opportunities to estimate the effects on the broader field of practice through internally maintained client databases. A hypothetical example would be finding out the number of teachers trained in a particular set of teaching techniques developed by NSF programs and incorporated in National Diffusion Network physics programs. This is a simple matter in the case of the NDN, because information about teachers trained by specific programs is a data element maintained in a central database.

- Finally, agencies with technical assistance functions are of special interest because they generally assume a proactive role that increases the likelihood of contact with NSF-generated knowledge. That proactivity is manifested in the "scanning" associated with technical assistance agencies; by design, they are searching continuously for emerging issues and perspectives within a number of environments. Further, technical assistance interests draw upon the varied worlds of research, policy making, and education practice. Thus, technical assistance intermediaries are likely to find useful a wider variety of types of knowledge generated by NSF programs than other agencies that may be interested exclusively in direct use training materials or research to shape policy.

Developing a Perspective: Influences on the Approach

The Footprint Metaphor- The metaphor of the footprint is a helpful starting point for thinking about reasonable boundaries for an evaluation, the nature of evaluation questions, and some options for data collection. "Footprint" signifies a mark or effect that will remain visible, at least for a certain time period. The footprint metaphor also suggests an evaluation that is concerned about what marks are made, how marks are made, and where they can or should be found. The metaphor suggests that the impressions left by an NSF program may be of varied depths, more or less visible, and more or less lasting. Much of the variation in impressions has to do with the other part of the metaphor: the surfaces on which the footprints fall. The approach in this paper emphasizes looking for the most appropriate (and one might argue, the most important) surfaces among the candidate intermediaries, meaning those that are most likely to accept, hold, and then even preserve footprints. The surfaces proposed here are examples from the national, regional, and state agencies or interest groups that have educational dissemination and reform support functions. In Karen Seashore Louis' (1981) terms, they are agencies that have external agent functions and multiple roles related to knowledge utilization: decision making, enlightenment, and capacity building. The enlightenment function (Weiss, 1972) of providing information and using research and development knowledge is especially relevant to the roles played by intermediaries as links in the research into practice continuum and to the type of knowledge generated by NSF programs. Technical assistance missions

"The knowledge utilization literature also points to the importance of the characteristics of what knowledge gets used and the conditions surrounding use."

suggest that the relationship between the intermediaries and educational practitioners is ongoing, characterized by gradual infusion of improved information and gradual learning and change.

Relevant Lessons from Research - The lessons from knowledge utilization and implementation research that may be most pertinent to this evaluation are cautions about what not to expect for outcomes, heightening sensitivities about what would be of value and interest, where to look, and what to expect. First, it is useful to review a few lessons from Milbray McLaughlin's influential summary (1987) of two decades of implementation research:

- We know to expect enormous variations in how knowledge is used, even when the object at hand is as bounded and prescribed as a packaged curriculum;
- We know that local capacity, motivation, and beliefs are the central influences on what gets implemented; and
- We know that it is individuals within organizations who use information, reflect on attitudes, and implement changes (not the organizations as units).

These lessons suggest modest expectations for knowledge use by intermediaries and practitioners. At the same time, McLaughlin's lessons suggest we need to ask what kinds of choices, interpretations, and transformations are made to meet the information needs of different actors at different points in time. They offer intriguing possibilities for questions about the capacities, motivations, and decisions that face intermediaries as they select and shape knowledge to influence practitioners. McLaughlin's "implementing system" notion suggests attention to the connection between the knowledge generated by NSF programs and those most likely to seek and make important use of it.

Research about the utilization of social science information offers other relevant lessons to frame the questions to be answered by an evaluation:

- Since utilization of knowledge takes many forms, and is seldom used in direct instrumental ways, the relevant questions related to use are when, under what circumstances, and how (Nelson, 1987).
- When viewed from a communications perspective, the important variables related to use are source, message, channel (the path and form of information), characteristics of the receiver, and conceptual impact (as opposed to instrumental) (Nelson, 1987).
- Utilization value depends partly on strategic conditions — timing, feasibility, values, and power orientation (van de Vall, 1987).

The knowledge utilization literature also points to the importance of the characteristics of what knowledge gets used and the conditions surrounding use. The variety of conditions surrounding the paths of knowledge use traceable to intermediaries is great. The ideal result from this type of evaluation is a deeper understanding from selected cases of how knowledge comes to be valued and used by intermediaries.

Context: The Author's Perspective- The design choices proposed in this paper about what would be both interesting and important to evaluate have been strongly influenced by my own work as a technical assistor in national educational dissemination and reform efforts.

In large part, my work and that of my RMC Research colleagues has been about support for reform of practice in teaching and learning at state and local levels, usually functioning as the type of dissemination/reform intermediary described in this paper. Our work for Federal and state governments and foundations is about the promotion of research-based policies and practices through training, consultation, and product development. As a group, we serve in several capacities as intermediaries, translating research into practice and supporting or facilitating educational improvements that contribute to systemic reforms. These responsibilities have directly impressed upon me a respect for the challenges involved in "leaving a mark" of any type on practice — even when one understands the complexities involved in influencing changes in educators' behaviors and attitudes, and is immersed in the policies and procedures of school systems.

At the same time, it is also clear that desired reforms do occur in some situations and under certain circumstances. And it is also clear that intermediaries have played a variety of roles in the reform process: stimulating dialog, providing background information, planning evaluations, interpreting results, working in partnership with schools to identify and implement changes, creating experiences to force disequilibrium, training teachers, linking with model programs, etc. My own experience has raised interest in (a) the proactive roles that the actors within intermediate agencies play in the transformation and transfer of knowledge into practice for the purposes of reform, especially now that the reform talk has turned systemic; (b) the processes by which we intermediaries shape, renew, and revamp our own knowledge bases; and (c) the group and individual decision making within intermediaries for selecting and sharing knowledge with practitioners.

Conceptual Framework: The Questions Addressed by the Evaluation

The conceptual framework for the evaluation design begins with a "macro" view of how knowledge affects practice. Exhibit 1 is a preliminary conceptual framework, illustrating components of a model with the following characteristics:

- Within the array of NSF grantees, specific elements of program-generated knowledge will need to be selected and described for tracking purposes;

- The paths of knowledge acquisition and transfer can be simple or multiple, circuitous or direct, connected or unconnected, curious, unpredictable, serendipitous, mutual — and are best traced through exploratory, investigative activity; the path-arrows on the diagram are meant to illustrate the wide variety of patterns that might be found;

- Intermediaries vary in scope, importance, function, and role;

- Intermediaries seek knowledge from and are influenced by many sources, including NSF programs;

- Intermediaries use a variety of modes to influence educational practice; and

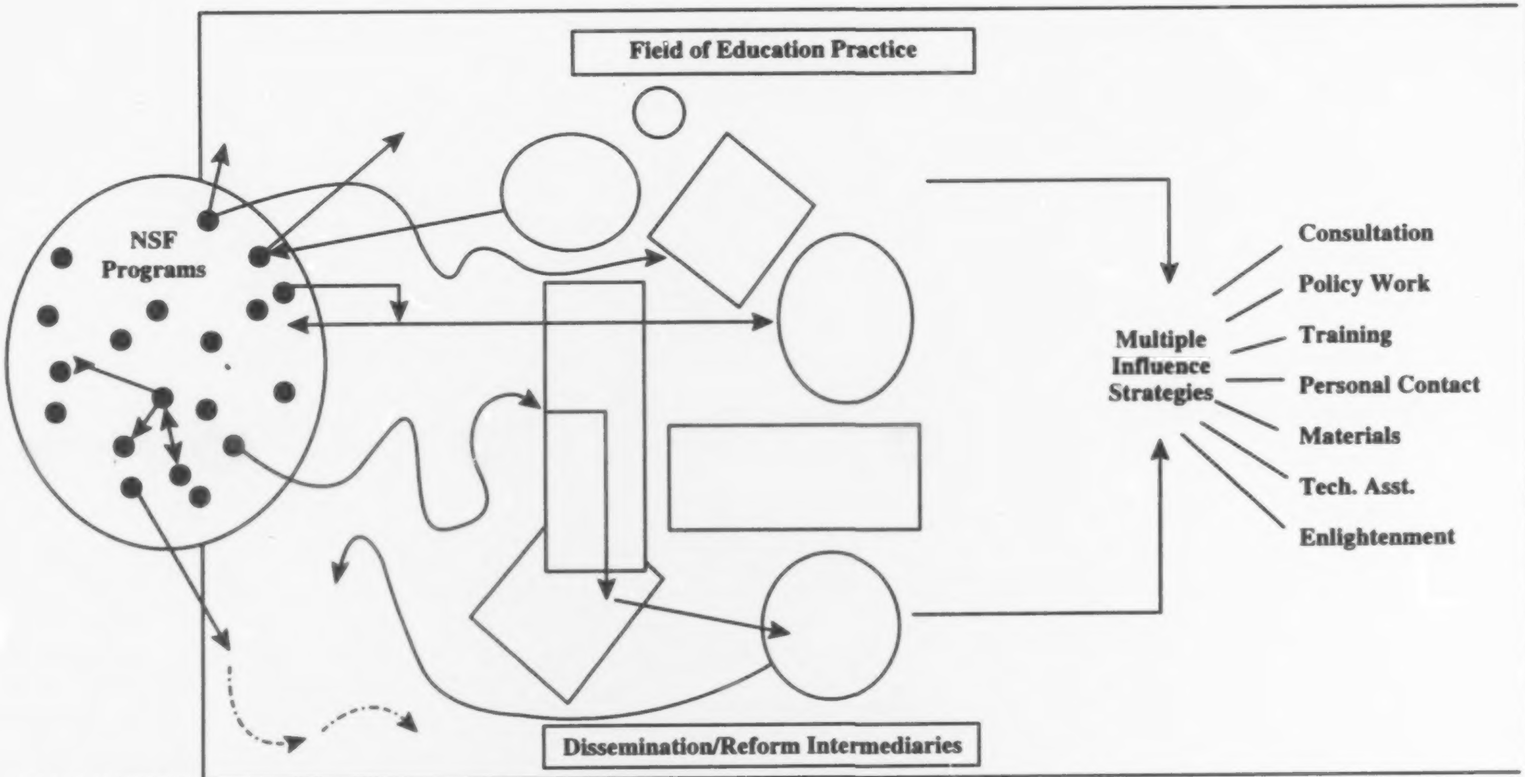
- NSF programs and the intermediaries are depicted within the field of education practice; obviously, both are also influenced by other elements of the field (although this is not depicted in the diagram simply to keep the discussion simpler).

As with any diagram of this type, this framework risks making the relationships among components seem less complex than they really are, but it does help to generate evaluation questions.

Evaluation Questions- At the simplest level, the basic evaluation question is about the very existence of footprints: Do knowledge footprints associated with NSF programs appear when the work and operations of intermediaries are examined? While practically challenging, whether the marks are found or not, this question is unlikely to yield information that is helpful to the ultimate purpose of NSF programs; that is, building a knowledge base that contributes importantly to reform of educational practice. Rather, the most interesting and useful questions involve asking where the footprints appear and about how they got there; what varied paths the footprints have taken; and the shape, size, and depth of the marks when located. These questions are important because dissemination paths were not originally prescribed, and therefore grantees' intentions about knowledge use are likely to vary dramatically. The utility of this evaluation is learning how knowledge reaches intermediaries; how intermediaries understand, select, and transform that knowledge; and how and to whom intermediaries promote the results. The special feature of the proposed approach is tracing both forward and backward; that is, following the paths of influences both in those cases where NSF program grantees intended dissemination for particular uses and in those where no proactive dissemination was intended.

Exhibit 1 lists four broad evaluation questions, corresponding to the relationships and processes depicted. The first two questions (What is the nature of knowledge that reaches intermediaries? and What are the paths and processes of acquisition and transfer?) seem most relevant. The question of how intermediaries translate and shape knowledge occurs at a different stage of the system of influence and is probably beyond the scope of this evaluation. The fourth question (How is knowledge used by intermediaries?) should be addressed to the extent of learning about intermediaries' intended uses of knowledge and their proposed strategies for influencing use. A beginning list of variables associated with the three questions (I, II and IV) of primary interest follows. Obviously, it would be important to involve stakeholders in identification and selection of the variables.

Exhibit 1. Conceptual Framework For Evaluation and Abbreviated Evaluation Questions


Field of Education Practice
NSF Programs
Multiple Influence Strategies
Consultation
Policy Work
Training
Personal Contact
Materials
Tech. Asst.
Enlightenment
Dissemination/Reform Intermediaries

EVALUATION QUESTIONS	I	II	III	IV
	What is the nature of knowledge that reaches intermediaries?	What are the paths and processes of acquisition and transfer?	What are the processes of translation and transformation?	How is knowledge used by intermediaries? Who/what is object of influence?

I. What is the nature of NSF program-generated knowledge that reaches intermediaries? What is the nature of the knowledge that does not? What are the differences?

Sample variables:

- Scope of implied change/impact
- Perceived proximity to typical practice
- Perceived and actual technical difficulty of application
- Perceived and actual implementation difficulties
- Perceived and actual degree of innovativeness
- Length of time available
- Producing institution and its affiliations
- Content
- Level
- Form, i.e., degree of "packaging" for practice
- Variety of channels and opportunities through which knowledge is available
- Amount of time investment required for initial understanding
- Directness of connection to national/state policies
- Directness of connection to student outcomes

II. What paths and processes do intermediaries use to acquire and receive NSF program-generated knowledge?

Sample variables

- Motivations and purposes for transfer
- Motivations and purposes for acquisition
- Direction of initiation
- Characteristics of initiators
- Roles and positions of key actors
- Formal relationships that facilitate transfer
- Forums for awareness and exchange
- Roles of professional associations
- Roles of colleges and universities
- Differences in initial and subsequent contacts with knowledge
- Similarity/difference with other acquisition activities of intermediaries, especially those related to mathematics, science, technology
- Barriers (attitudes, understanding) from multiple perspectives
- Role and context of personal contact
- Function of receiving unit within intermediary
- Perceived satisfaction
- Content expertise of receiver

IV. How is NSF program-generated knowledge used by intermediaries? What are the intended uses and strategies that connect to education practice?

Sample variables:

- Internal and external enlightenment functions

- Basic modes/strategies of influence
- Similarity of mode to typical strategies
- Placement within ongoing functions
- Fit within ongoing conceptual work
- Fit within ongoing instrumental work
- Stimulus for new approach/new activity/revamping
- Facilitation of connection to different levels of practitioners
- Perceived satisfaction
- Additional needs associated with intended uses

Study Method- Case histories are a logical data collection strategy, based on multimethod exploratory investigations that trace (a) forward from knowledge examples provided in NSF projects and (b) backward from selected intermediaries. Several case histories would yield a detailed picture of some of the effects of knowledge on intermediaries, and by extension, the effects on education practice. Similarly, the study would also identify knowledge that did not reach intermediaries. Judgments about the value of the emergent patterns of knowledge use and non-use become a stakeholder problem, but one that might be resolved through other parts of the evaluation—perhaps, for example, the independent expert assessments of the value of NSF project work that were suggested by several other paper authors. Cross-case analyses (using the intermediary as the unit of analysis) would provide information about how knowledge is or is not acquired and transferred.

An Extended Example and Some Practical Problems

We have not yet addressed the scale of the evaluation. A modest but indepth exploration of three to four well-selected intermediaries would be sufficient to (a) learn about the value of the approach and (b) gather enough leads about influences on intermediaries to preview effects. Obviously, the selection of intermediaries is critical; consensus on their representativeness, potential for depth and breadth of contact with NSF program-generated knowledge, and effectiveness in technical assistance and reform must be established among stakeholders early in the evaluation. The intermediaries should probably represent a wide range in terms of likelihood of use of NSF program-generated knowledge, ranging from obvious choices (i.e., those with direct and primary roles in the reform of mathematics and science education practice and the application of technologies) to those with strong and important connections to practice but less obvious connections to NSF programs.

> *"... the study would also identify knowledge that did not reach intermediaries."*

The extended example that follows is an unlikely intermediary, chosen to illustrate the potential of the approach to uncover effects. The example previews the issues that will arise in identifying and selecting candidates and collecting data. The sample intermediary is the national network of Chapter 1 Technical Assistance Centers (TACs) and Rural Technical Assistance Centers (R-TACs). Its selection was based on the author's experience with TAC operations and not because it necessarily represents an optimal candidate. TACs are unlikely intermediaries because they are not charged with reform of mathematics and science teaching.

"TACs connect knowledge and research to practice in mathematics to an extent far greater than a passing acquaintance with the TAC network might suggest."

The TAC network comprises 16 Federally supported multipurpose centers (approximately 65-70 full time equivalent) serving state and local education agencies in the areas of Chapter 1 program design and improvement and program assessment. In the past 5 years, TAC activities at the local level have focused largely on improvements in Chapter 1 programs, including promotion of research-based strategies for teaching and learning in mathematics, reading, and writing. The ultimate beneficiaries of Chapter 1, and therefore TAC activity, are disadvantaged students and their parents at all levels. In elementary mathematics, for example, TAC activities might well include any or all of the following:

- Identifying curriculum and materials;
- Providing awareness of the National Council of Teachers of Mathematics (NCTM) standards directly to practitioners;
- Consulting with districts to establish staff development structures;
- Training administrators, teachers, and curriculum specialists in research-based principles, strategies, and techniques;
- Conducting demonstration lessons as part of inservice work;
- Helping to locate or design alternative assessments for problem solving;
- Researching the practices of other states regarding criteria associated with standards;
- Introducing parent leaders to the principles associated with advanced skills in mathematics, and defining high expectations by example;
- Developing research syntheses to inform policy development at the state level;
- Representing compensatory education interests on a statewide committee interested in reform;
- Encouraging a district to address the weaknesses of mathematics curriculum when developing a program improvement plan;
- Helping districts interpret the implications for instruction of the results of mathematics assessments;
- Writing a newsletter article on high-powered strategies for disadvantaged learners;
- Gathering information about user experiences with particular software; and
- Developing an agenda and locating presenters for a regional or national conference on mathematics teaching strategies for disadvantaged learners.

Certainly, TACs are not the only resource that Chapter 1 programs turn to for support in reform of the teaching of mathematics. However, because the TACs are multipurpose, credible, and provide services at no cost, there is a tendency for Chapter 1 clients to contact them for a wide variety of functions, as the above list demonstrates. As a result, TACs connect knowledge and research to practice in mathematics to an extent far greater than a passing acquaintance with the TAC network might suggest. Furthermore, they are working with practitioners who serve an especially important group of students—students who are economically and educationally disadvantaged.

TACs have several other features that raise issues about what makes a good candidate to be an intermediary:

- Only a small proportion of TAC staff (perhaps 10 percent) have academic backgrounds related directly to elementary mathematics, so the need for acquisition of knowledge to serve clients is clear;

- Materials and training are shared across the network through established mechanisms (quarterly meetings, institutes and seminars, an electronic network, materials clearinghouses, some common policies related to materials development, a culture that supports exchange) so influences can spread fairly rapidly; and

- Two separately funded support centers for the TACs have the mission of acquiring knowledge related to curriculum and instruction and organizing, translating, transforming, and disseminating it for use by all the TACS.

The question of whether or not the TAC network would be a viable candidate for intermediary status in this evaluation can probably be answered by the degree to which the reader is intrigued at this point to find out how TACs have been using the knowledge generated by NSF programs. Intermediary selection criteria emerge from consideration of advantages and disadvantages of the TACs as candidates for this study.

The advantages are the national scope of TAC influence; the mission of improving educational programs for the disadvantaged at all levels; the multiple functions of training, policy support, planning, consultation, and product design; the simultaneous work at different levels of educational practice (classroom, school, program, district, state, regional, and national); some degree of commitment for generating improved knowledge for practitioners; the system support for enlightenment and capacity-building uses of knowledge; the relatively small size of the network and accessibility of personnel; and the capability of tracking activities through content-based client service records. The disadvantages are the multipurpose TAC mission; competing obligations, because the TAC agendas are determined largely at state and Federal levels; and the variability of knowledge use across TAC centers (as a result of organizational context and cultures as well as regional needs and interests). These advantages and disadvantages offer a preview of criteria that might be used to select intermediaries for study.

Speculating on the results of an exploratory review of NSF influences on TACs leads to these hypotheses: the influences would be numerous; TACs would probably be the initiators of knowledge acquisition, using some traditional awareness vehicles but often becoming aware of specific knowledge through policy-related channels (Federal policy studies, for example); the primary intention of knowledge use by TACs would be teacher training through their influence on program design and policy development; knowledge with the clearest connection to student outcomes would be preferred; TACs would expect to translate research findings into best practices before using them with practitioners, even for enlightenment functions; and TACs would have a strong interest in assessments and perhaps initial contact would have been based on interest in assessment.

"... the evaluation process is more like an investigative dialog with intermediaries than a survey of use ..."

Selecting Intermediaries- The extended TAC example of an unlikely intermediary candidate raises interest in thinking about intermediaries that would be viable. The example also introduces a host of practical problems to be faced in the study, beginning with the process of identifying and selecting intermediaries. Different stakeholders will have preferences for different types of intermediaries. The essential criteria for intermediaries should be potential and credibility—potential in terms of likelihood of locating effects and credibility in the sense of scope and importance of influence. Other related criteria are national profile and scope of influence; longevity and stability of the intermediary; clarity of mission with respect to dissemination and reform; proactivity of outreach and extent of collaboration with other intermediaries; multiple functions, including a research and development capacity; multiple entry points from the perspective of practitioners; and maintenance of records that permit tracking of client contact at some level.

A related issue will be identifying the best informants or reporters from each intermediary, recognizing that the functions of knowledge acquisition, transformation, and use are probably carried out by different units within an intermediary.

Bounding Data Collection- This is perhaps the most elusive element of the proposed approach. There is little guidance for knowing what program-generated knowledge would be best for tracking purposes. Selecting and defining knowledge would involve those individuals or groups most familiar with the knowledge generated by NSF projects, especially the grantees themselves. Grantees are in the best position to know the aspects of their work that have potential for influencing practice and to identify what has already found a way into practice. We envision a process that engages grantees and other stakeholders (e.g., NSF, selected intermediaries) in developing a set of theories about the presumed paths of influence associated with knowledge generated from their work. That process would yield a range of types of knowledge to be developed into descriptions for tracking purposes.

Because a key evaluation purpose is to learn primarily about the process of acquiring, using, and valuing knowledge, it would be important to select examples of knowledge that are concrete, as well as examples that would be more difficult to track. Ideally, the pool of descriptions would vary at the outset and from 0 perspective by format, scope, content and level, proximity to practice, longevity, perceived innovativeness, and technical complexity. Because the evaluation process is more like an investigative dialog with intermediaries than a survey of use, descriptions need only serve as conversation starters, not complete catalogs of program-generated knowledge. An obvious challenge is that intermediaries will have translated and transformed the knowledge as they have incorporated it into their work.

Data Collection Procedures and Analysis- The exploratory nature of tracking the influence of knowledge suggests use of the investigative journalism metaphors and models described by Smith (1981) and Guba (1981). In Guba's terms, the goal of tracking the paths of influence is to develop "working hypotheses embedded in thick descriptions." Evaluators follow the trail of a chain of events, continuously using creative strategies to develop new sources and leads. The process requires the establishment of a record, reconstructing and then verifying the tracks. Next steps always proceed from what has been previously documented, analyzed, and summarized. Continual recycling to previous sources and leads with newly generated hypotheses is part of the process, as is running information back through contacts for confirmation or refutation. Data collection includes records review and analysis, key interviews, and observations to "establish a record" of transactions, profiles, chronologies, and relationships. Developing and refining hypotheses about what and how influence occurs is a matter of cross-referencing the varied pieces of information in the rich database built from the experiences of the intermediaries and the points of contact with NSF projects. As data are collected, the conceptual framework would be refined through reflection on the evolving hypotheses. Finally, cross-case analysis (each intermediary is a case) would be based on a revised framework. Both cross-case results and descriptions of the experiences of each intermediary represent valuable products.

Alternatively, one might organize and vary data collection by stages, beginning with surveys and/or focus groups of grantees to learn first about possible intermediaries, presumed paths of influence, and dissemination intentions. Next, paths might be traced backwards from intermediaries as discussed above. Grantees with especially clear, deep, and varied connections to knowledge could then be studied in depth in another set of cases.

Summary- Suggestions provided by authors of other papers in the series offer potential solutions for methodological issues that arise from this evaluation approach. Cluster evaluation techniques could be used to define knowledge and hypothesize paths to trace. Expert judgments are required at several points to give value to the findings about paths and processes. Generalizability analysis as discussed by Webb offers some ideas for sampling the breadth of NSF programs. Several elements of Yin's partial comparison model provide conceptual guidance for thinking about the legitimacy of approaches: use of proximal outcomes where interventions are weak, direct assessment of process logic, and the value of compelling explanations of documentable chains of events.

The connections to other papers reinforce the possibilities for using this evaluation approach in conjunction with others. The approach is offered as one of several "small wins" in Karl Weick's terms—an option for breaking down the complex nontraditional evaluation problem into a series of achievable tasks.

References

Barley, Z.A., and Jenness, M. 1993. *Conceptual underpinnings for program evaluation of major public importance: Collaborative stakeholder involvement.*

Bickman, L., ed. 1987. *Using program theory in evaluation.* San Francisco: Jossey-Bass.

Bryk, A.S., and Hermanson, K.L. 1993. Educational indicator systems: Observations on their structure, interpretation, and use. In *Review of Research in Education*, vol. 19. Washington, D.C.: AERA.

Guba, E.G. 1981. Investigative journalism. In *New techniques for evaluation*, N.L. Smith, (ed.). Beverly Hills, CA: Sage Publications.

Knapp, M.S., Shields, P.M., and Turnbull, B.J. 1992. *Academic challenge for the children of poverty.* Washington, DC.: U.S. Department of Education.

Knapp, M.S., and Turnbull, B.J. 1991. *Better schooling for the children of poverty: Alternatives to conventional wisdom.* Berkeley, CA: McCutchan.

Knapp, M.S., Zucker, A.A., Adelman, N.E., and St. John, M. 1991. *The Eisenhower Mathematics and Science Education Program: An enabling resource for reform.* Washington, DC.: U.S. Department of Education.

Little, J.W. (Forthcoming). Teacher's professional development in a climate of educational reform. *Educational Evaluation and Policy Analysis.*

Louis, K.S. 1981. External agents and knowledge utilization. In *Improving schools: Using what we know*, R. Lehming and M. Kane (eds.). Beverly Hills, CA: Sage Publications.

McKinney, Kay. 1993. *Improving math and science teaching.* Washington, D.C: GPO.

McLaughlin, M.W. 1987. Learning from experience: Lessons from policy implementation. *Educational Evaluation and Policy Analysis* 9 (2).

Nelson, C.E., Roberts, J., Maederer, C., Wertheimer, B., and Johnson, B. 1987. The Utilization of social science information by policymakers. In *American Behavioral Scientist,* 30 (6), Sage Publications.

Smith, N.L., ed. 1981. *Federal efforts to develop new evaluation methods.* San Francisco: Jossey- Bass.

Smith, N.L., ed. 1982. *Field assessments of innovative evaluation methods.* San Francisco: Jossey-Bass.

Smith, N.L., ed. 1981. *New techniques for evaluation.* Beverly Hills, CA: Sage Publications.

Suarez, T.M., Montgomery, D.L., and Dania, S.T. 1989. Stakeholder II: The use of stakeholder analyses of anticipated outcomes in the design of summative evaluation. Paper presented at the meeting of the American Evaluation Association, San Francisco.

van de Vall, M. 1987. Data-based sociological practice. In *American Behavioral Scientist,* 30 (6), Sage Publications.

Weick, K. 1984. Small wins: Redefining the scale of social problems. *American Psychologist* 39, 40-61.

Weiss, C. 1972. Evaluation research. Englewood Cliffs, NJ: Prentice-Hall.

Wicker, A.W. 1985. Getting out of our conceptual ruts: Strategies for expanding conceptual frameworks. *American Psychologist,* 40 (10).

Yin, R.K. 1993. *Evaluation design: Breaking new ground.*

Robert Mac West - RMW Science Action

As I listen to these papers and discussions, it becomes quite clear that I represent another face of science education—that which calls itself informal science education and includes institutions such as natural history museums, science centers, children's museums, zoos, aquariums, botanical gardens, community centers, youth organizations (4-H, Girls Inc.), and even theme parks. Informal science education generally is not well connected to the educational research loop. However, (consciously or unconsciously) it both uses the results of educational research and seeks to avoid them. Let me make some connections and observations.

- Because informal science often is an open system—museum visitors (apart from highly directed school visitation) use museums as a recreational outlet as much or more than as an educational experience—museums have to look at the educational process differently from the way the formal system does. This has forced museums (or at least the currently more successful ones) into serious front-end evaluation, or needs assessment. They are customer-driven, rather than driven by current research or even the availability of technology as delivery system, even though "research says" that museums must devise some other strategy for presenting that particular body of material. If a museum goes ahead simply on the basis of research studies and staff-initiated approaches, it runs the risk of giving a party to which no one comes. Therefore, museums are very selective about what they use of the enormous body of research that is being generated by NSF, intermediaries, and other providers of materials and ideas.

- Also, because most museum visitors are self-selected and not part of any curriculum or content module, most museums have learned that they must be very careful about defining desired outcomes. Measurements of content knowledge are unimportant to dangerous. The expectation that one several-hour visit to a science center or natural history museum will make or break a scientific career is preposterous. Rather, as in Hezel's discussion of goal-free or naturalistic evaluation, they are concerned with attitudes, and like Mark St. John, have become very good at ferreting out reactions to science as a way of thought and a legitimate area of interest and learning.

- Museums are experimenting with new techniques in data gathering. Webb mentions videotaping as a way of recording events and experiences. Museums do a great deal of that, sometimes even recognizing at the time filming is ongoing that the presence of a camera may cause behaviors and reactions that would be quite different were the camera not there. They do a great deal of eavesdropping, both surreptitiously and openly, following visitors to see what they do. In these respects, museums are able to be more creative than are researchers attempting to understand what goes on in a classroom.

- However, there are times when museums behave like the formal system and even are integrated into schools, and when museums serve as Dwyer's intermediaries for dissemination of ideas and materials into schools. For many years the Lawrence Hall of Science, a unit of California-Berkeley, has produced curriculum materials. These are marketed as GEMS—Great Explorations in Math and Science. These materials are tested extensively, and teacher workshops are convened to assist with their penetration of the classroom. A number of projects funded by the Howard Hughes Medical Institute are generating more materials tightly connected to reform curricu-

> la—the University of Nebraska State Museum, the Oklahoma Museum of Natural History, and the Buffalo Museum of Science are deeply involved in this effort.

In conclusion, I suggest that there is a science education universe that calls itself informal science education. Sometimes it is responsive to the varied research being done through NSF, mainly when it sees that there is a clear utility to the NSF products. And just as often, this universe sees the research efforts as being unimportant or producing inapplicable results. Because museums are a growth industry, and because they are becoming increasingly sophisticated at knowing what happens in their exhibits and programs, it may be useful for evaluation of NSF research efforts to begin to include their impact on the informal sector.

I must say that I am very tempted by the previous presentations to tell you about what our Center does, but since I was always a good little student who did what the teacher said, instead I will do as asked and comment on the three papers I was asked to review. First of all, I too found them quite eclectic and interesting. I'll give you some general comments and then try to be specific.

One paper begins with the author characterizing the program he is considering. I found this very useful because it sets a context that I thought was needed for the suggestions and recommendations he makes, even though these may be quite generalizable. Two of the papers make specific evaluation suggestions, some of which were presented orally by the authors. The third paper was quite general. Two of the papers (not the same two) give very long lists of particular evaluation questions that might be asked. As the last speaker pointed out, there already is a tremendous problem in bounding the questions, let alone the data collection efforts. The two papers giving long lists did not deal successfully with this problem.

From my perspective—and I've been in and out of government—all the papers were written very much from an evaluator's point of view rather than from the point of view of the clients for the evaluation. The clients for NSF are researchers and educational systems and institutions; those who ultimately are going to ask evaluation questions are OMB and Congress. If NSF doesn't satisfy these oversight bodies, none of us will be sitting here 5 years from now.

A common theme advocated by these papers is the need for dissemination. That raises the question of what is worth disseminating and how this is decided. Webb's paper discusses an internal process of self-evaluation. He suggests using videos as a way of communicating, but does not address the issue of the researcher needing to decide what he or she wants to communicate. In fact, none of the papers addresses this issue in any detail.

Dr. Hezel made a point (quoted from someone else) that is well worth talking about: The kind of evaluation we are considering at this conference should be conceived as part of a system of self-renewal rather than as a yes/no decision-making paradigm. I commend all the papers because I think they are written in that spirit; it is a critically important point of view in considering alternative nonaccountability types of evaluation.

A second major point that he makes in his paper concerns the importance of dissemination. But I see very little in this paper that tells me an acceptable way of deciding how or why one would want to disseminate particular evaluation findings. Dissemination costs lots of money; let's not fool ourselves about this. I've always been amused at the funding curve that characterizes the Federal government and also private foundations that support research and development in education, as contrasted to that of private industry. For education, research and development receive the lion's share of funds, followed by program development, and trailed by dissemination or marketing. Because Federal agencies haven't learned that dissemination is very costly, the issue of what one chooses to disseminate has to be taken seriously.

Several minor points about the Hezel paper. He urges against nose counting, which I appreciate in the context of nontraditional evaluation. On the other hand, he also rightly points out that one must count whether minority populations, disadvantaged populations, and so on, are getting benefits. Therefore, I found myself in a little bit of a quandary as to whether the paper advocates nose counting or warns against it. This needs clarification. There also is a distinction, made early in the paper, between qualitative types of methods as being appropriate for non-traditional evaluations and quantitative types of methods being appropriate for the usual sorts of outcome and impact assessments. In fact, later in the paper the author suggests using both quantitative and qualitative methods to address evaluation questions of both types. I agree with his later statement; the author should reconcile these two apparently conflicting positions.

Hezel also discusses the notion of tracing the intellectual origins of an innovation or program being evaluated. It would be extremely difficult for me to do so in my own work. I can't document how my synapses work. Yes, as researchers and evaluators, we add long lists of citations at the end of everything we write, whether it's a proposal or a paper, but where the intellectual ideas actually came from and how they were synthesized to give rise to a new project would not be easy to trace. Another question Hezel suggests asking concerns duration of the project and the difference between the proposed and actual duration of the project. I'm not sure what that would tell us, but perhaps Dr. Hezel could respond.

The next paper I want to comment on is Chris Dwyer's. (In my comments, I am moving from the most general to the most specific paper.) She discusses the idea of using intermediaries as key informants. This is a useful approach, provided the intermediaries are chosen appropriately. In my view, however, her list of criteria on selecting the intermediaries is missing the most important one, that is, whether the intermediary is knowledge-searching. Does it even operate in a context in which it needs R&D and evaluation knowledge? If so, what are its search mechanisms and the filters it uses for selecting what to act on? A good example is one that Dwyer actually gives, the National Diffusion Network (NDN), which uses a very particular kind of filtering device for deciding whether to disseminate information about a given program or not. If the desired evaluation (or filter) wasn't built into the program in the first place, it will never make it through NDN because it doesn't provide the data on which NDN bases its decision. Or to put it differently, NDN defines quality through impact data, while the program may have quite different criteria. Another question concerns how the intermediary deals with the information it acquires. If the intermediary is a knowledge-seeking kind of organization, if it has defined filters by which it judges the quality of research reports, research and development products, or whatever, and if it also has a process for acting on its searches and judgments, then I agree that including intermediaries as key informants is one strategy among a number that could be considered.

I would not start in the way she suggests, however; I would do some retrospective analyses, namely look at the intermediaries and what knowledge they are actually using. That may raise similar problems to those I noted earlier with respect to Dr. Hezel's recommendation on tracing the origins of ideas. Perhaps one could start with some specific practice that looks as if it had been influenced by some assessed program, and then trace back where the practice came from. If the tracing involves an intermediary, the practice may have multiple origins. A good intermediary, one that is out there to improve practice, should be using multiple sources of information, not merely relying on a single project or program as the sole source for its information.

I found Norm Webb's paper very interesting and thoughtful. He placed his discussion in the context of a specific program, so that one could follow how he was relating his four major suggestions to NSF's Research on Teaching and Learning (RTL) Program. The evaluation matrix he suggests makes us aware of having to look for both the successes and the failures. Failure contributes to our knowledge as well as success; we tend to forget that. We tend to believe that only success is good, but that's not true in research or even in development. For example, we may develop a program that works in some setting, but when we find out it doesn't work in other settings, that's very important information. Webb's matrix reminds us of this.

Let me comment on his specific suggestions. Regarding the retrograde analyses, I may have misunderstood what he intends, but I think they might focus too much on the internal process of a particular researcher or project. I would feel that's too narrow a net, unless combined with other strategies. If it is just one component of an evaluation, then I think it's an interesting suggestion. Something like that might be a piece of a larger-scale evaluation of an NSF program.

This particular suggestion reinforces the general impression I had of all the papers I reviewed, namely, that they appear to be written from an evaluation rather than from a policy perspective. For example, Webb conjectures about the reason for the many and varied kinds of projects in the RTL Program. Possibly, as he says, this has to do with all the client audiences, their needs, and all the different avenues to pursue. More likely, since this is a field-initiated program, and the peer review system being what it is, I suspect the eclectic nature of the RTL Program comes about as much through proposal pressure exerted by good people proposing the things they want to do as through a desire to meet client needs. The perspective of the evaluator of education R&D is different; we are concerned with the use of R&D products. So that's why I say this set of papers is written from the perspective of the evaluator rather than the real world of Federal agencies, but that's fine. I commend NSF for going outside its own concerns to get a different sort of perspective.

I've noted that I feel Webb's first suggestion is too narrow—just looking at NSF generating its own further work through its principal investigators. The second suggestion, video documentation, made more sense to me in his oral presentation than when I read it. When I read it, it seemed more like PR than like evaluation. But orally, Webb made the point that, in the process of creating such a video, one would have to think about what it is that is important to disseminate. I think that's a very valid point, as I noted earlier. But I want to reemphasize that there has to be more widespread dissemination than just to one's peers; that is, to people who generate the research or the development products or who make judgments about what is worthy of dissemination.

The suggestion for cultural analysis of the research community is a wonderful idea, but I wonder whether it will be of interest to Congress. Consider the creation of a community of scholars that can engage in the kind of dialogue we are having this morning. This seems like a good thing. However, I am reminded of something that happened in the 70s when lots of money was being poured into graduate fellowships and traineeships in order to create a science infrastructure. All of a sudden, there were lots of young researchers asking for research money, and OMB said "Oh, we have created a monster. We cloned all these researchers and now we've got to feed them. This has to stop." And it did stop. Well, all right, I love the idea of studying the research community, but history makes me ask, "What is the Hill going to say?"

The fourth suggestion that Webb makes is on generalizability analysis. I have not had a chance to see the paper from Western Michigan, so I'm not precisely sure what it says about cluster analysis. This may be a better approach than a statistical one. Random sampling to deal with the great variety of projects funded by a program such as RTL does not strike me as appropriate. I would prefer groupings of projects that in some way reflect the approach taken, the problem addressed, etc. The groupings would have to be thought through very carefully. After grouping, one might select a representative subset of projects from each group for evaluation. If, for example, you had 200 projects and created 10 groups, you could select 3 out of each group for further study. I think that might be a better approach than random sampling.

Let me end my remarks by thanking NSF for the opportunity to participate in this stimulating conference and the audience for your attention. I look forward to the publication of all the papers.

Conceptual Underpinnings For Program Evaluations Of Major Public Importance: Collaborative Stakeholder Involvement

Zoe A. Barley and Mark Jenness
Western Michigan University

Overview

This paper suggests that three considerations should prevail in the evaluation of National Science Foundation (NSF) programs. First, evaluations of major public significance should provide for a process that gives voice to the key stakeholders of the evaluation. Second, evaluation should be designed and implemented to serve a primary function of program improvement, including enhancing dissemination. Third, NSF program evaluations should be exemplars for individual project evaluations.

Current issues in evaluation that have emerged from a need to develop program evaluations relevant to a wide variety of audiences (stakeholders) are briefly discussed. Additionally, emphasis is placed on the importance of using evaluations to shape programs to enhance effectiveness as they are in progress, rather than on providing post hoc findings that are often not amenable to real world adaptation (dissemination).

As a strategy for evaluation, this paper describes a method of evaluation of multiple projects with common or closely similar outcomes that has been named "cluster evaluation."[1] While aspects of the method can be used retrospectively and could be used to aggregate findings from a program's funded projects, a primary value of cluster evaluation is in the formation—during the course of program activity—of an interactive, collaborative group consisting of project directors, funding agency program staff, evaluators, and other appropriate key stakeholders. Cluster evaluation is, therefore, constructivist in orientation, with the evaluation being constructed out of the shared visions, values, and directions of the cluster group.

Cluster evaluation—a collaborative, project-enhancing, leadership-enabling, outcome and system-focused process—is an appropriate framework for any publicly significant evaluation. Certain elements of this evaluation method may be more directly relevant for overall evaluations of NSF programs than others, and the process can be adjusted to meet the needs of particular situations.

"... evaluations of major public significance should provide for a process that gives voice to the key stakeholders of the evaluation."

Implications for NSF Programs

In its efforts to identify nontraditional approaches to program evaluation (the "Footprints" project), NSF can learn much from cluster evaluation methodology and its philosophical underpinnings. Diverse multisite programs with common areas of interest seeking to improve overall and individual project efforts and determine effects of program process and accomplishment of outcomes, such as those of NSF, are primary candidates for cluster evaluation. Although cluster evaluation can be applied to many settings in and out of government, for the purposes of this paper, reference will be made to the NSF Research in Teaching and Learning (RTL) program as an example for applying cluster evaluation.

---

[1] This is a new use of the term "cluster evaluation" and bears no resemblance to evaluation forms existing prior to 1988.

"*The new evaluator is someone who believes in and is interested in helping programs and organizations succeed.*"

The RTL program, according to documents supplied to the authors as background for the "Footprints" assignment, "seeks to support new discoveries about how individuals and groups learn, teach, and work more effectively in complex, changing environments." Three important goals of the RTL program are particularly amenable to the use of cluster evaluation: 1) building a coherent and comprehensive base to meet future and current needs of all decision makers, 2) initiating an emphasis on direct teacher and other stakeholder involvement, and 3) helping assure the application of research findings. An interactive, collaborative evaluation process gives voice to front line educators, as well as researchers, in a nonthreatening, practical issues-focused context in which assessment and evaluation become tools to improve practice and to shape programs to serve the full range of interested audiences.

Statement of the Problem—Program Evaluations

Through its "Footprints" project, NSF is exploring alternative, nontraditional approaches to evaluating their efforts, especially those programs focusing on mathematics and science education research and applications of technology. Frechtling, in her introduction to the "Footprints" papers, discusses three concerns about traditional evaluations in the context of NSF needs: 1) given the multiplicity of influences, it is unlikely or impossible that appropriate unidimensional causal statements can be drawn, 2) sole use of quantitative measures are likely to exclude important information, and 3) impact measures, such as student achievement, need to be considered relative to the likelihood of impact in the projects' time frames.

These are important concerns and are discussed in more detail in the context of philosophical underpinnings of cluster evaluation. First, the nature of evaluation data needed, not only for funders but also for a wide array of audiences for the purpose of accountability, project refinement and enhancement, and successful dissemination, is much more complex than previously thought necessary and hence more difficult to obtain.

A second critical issue, however, lies in the purpose of the evaluation itself. Wholey (1983) saw parallels in the public sector use of evaluation to what profit does in the for-profit sector, providing critical feedback that is immediately useful to policymakers and managers. In his 1973 work (Wholey and White) he stated, "the main purpose for evaluation, . . . to feed back information about how a program is working to improve its operation, is missing from most local and state evaluation activities." In another article, he suggested,

> The new evaluator is a program advocate—not an advocate in the sense of an ideologue willing to manipulate data and to alter findings to secure next year's funding. The new evaluator is someone who believes in and is interested in helping programs and organizations succeed. At times the program advocate evaluator will play the traditional critic role; challenging basic program assumption, reporting lackluster performance, or identifying inefficiencies. The difference, however, is that criticism is not the end of performance-oriented evaluation; rather it is part of a

> larger process of program and organizational improvement (Bellavita, Wholey, and Abramson 1986, p. 289).

Finally, Frechtling notes recent trends in evaluation which seek to involve all the stakeholders in the process. Cronbach and associates (1980) see this as a key task in understanding the political nature of the end result of any important evaluation study. Guba and Lincoln (1989) speak of stakeholders' claims, concerns, and issues as organizers for the evaluation. Donmoyer (1991) strongly suggests that stakeholders be actively involved in dialogs before, during, and after the evaluation.

If these purposes and intents—implementing a more appropriately complex evaluation, shaping programming and improving projects by providing feedback during project implementation, and involving stakeholders in the process—are valid, the evaluations should be shaped "upfront" with these goals in mind. Grantees, however, are often not prepared with either the evaluation skills required or knowledge of the broader context in which their project findings are relevant for those findings to be meaningful. While some amount of information can be obtained from evaluations after the conclusion of projects, to achieve a measure of information appropriate for use in guiding selection of new projects, in disseminating results to other project sites, or for use in systemic change modalities, the evaluation process must be improved as the projects proceed.

Design Considerations—Undergirding Philosophy

Two conceptual models offer useful insights for designing nontraditional evaluations for NSF research-oriented programs: Cronbach's concept of a Social Problem Study Group from his 1980 book, *Toward Reform of Program Evaluation*, and Guba and Lincoln's fourth generation evaluation from the book (1989) by the same title. They also provide guidance in the design and implementation of cluster evaluation described in a later section.

Cronbach suggests the formation of a social problem study group made up of members representing all concerned parties for evaluations of social significance, not unlike panels NSF convenes for evaluation purposes. The group, however, would embrace the following activities:

- Study problems (e.g., What should be the influence and direction of an NSF program?) in the broadest possible way.

- Hear from those who conduct evaluations, preferably as their work progresses; hear from those who deal with the problem in service agencies; hear from those who have ideas about new policies and interventions.

- Produce a far more comprehensive and dependable interpretation than emerges from a single study or a lone critic questioning a finding.

- Continually reformulate the questions worth studying and recast key terms that define stated problems.

- Put research into proper time perspective, dispelling the illusion that quick and partial studies will resolve dilemmas.

- Provide a forum for putting observations and uncertainties into perspective.

- Be willing and able to think hard about the specified problems.

In another but related direction, Guba and Lincoln have defined "fourth generation evaluation" in which the processes of the evaluation are as follows:

1. Identifying the full array of stakeholders who are at risk in the projected evaluation.
2. Eliciting from each stakeholder group their constructions about the evaluation and the range of claims, concerns, and issues they wish to raise in relation to it.
3. Providing a context and a methodology through which different constructions, and different claims, concerns, and issues, can be understood, critiqued, and taken into account.
4. Generating consensus with respect to as many constructions, and their related claims, concerns, and issues, as possible.
5. Preparing an agenda for negotiation on items about which there is no, or incomplete, consensus.
6. Collecting and providing the information called for in the agenda for negotiation.
7. Establishing and mediating a forum of stakeholder representatives in which negotiation can take place.
8. Developing a report, probably several reports, that communicate to each stakeholder group any consensus on construction and any resolutions regarding the claims, concerns, and issues they have raised.
9. Recycling the evaluation once again to take up still unresolved constructions and their attendant claims, concerns, and issues (Guba and Lincoln 1989).

"Collaboratively, they generate an evaluation that is far more than monitoring or accountability, but which addresses broad-level policy considerations in a future-oriented mode."

Taken together these two frameworks suggest an evaluation process that actively involves all the known stakeholders. Collaboratively, they generate an evaluation that is far more than monitoring or accountability, but which addresses broad-level policy considerations in a future-oriented mode.

Evaluation Questions for NSF Programs

The following questions are suggested as the guiding overarching questions for evaluating NSF mathematics and science education programs, including the Research in Teaching and Learning program. Additional overarching questions and/or subquestions pertinent to a particular program area should be added as appropriate.

The use of concise questions in each of three areas—outcomes, context, and implementation—provides the perspective for not only reporting results, but also for understanding the conditions in which the results were obtained and the exact nature of the programming that produced the results, or lack thereof.

Outcome Questions

What has been the nature of the impact (intended and unintended) of the program on teachers and learners? Positive outcomes? Negative?

What has been the nature of the impact on the system of mathematics and science teaching and learning?

What kinds (and numbers) of new leadership have emerged within the educational system as a result of the program?

What new national or local programs and policies have emerged or been furthered as a result of the program?

Context Questions

Has the program effectively served a diverse body of mathematics and science educators?

Has the program effectively reached a broad range of mathematics and science learners?

For what educational settings has the program's effectiveness been demonstrated?

Has the program funded grantees across a broad range of characteristics representative of the educational system, especially in mathematics and science?

Implementation Questions

Has the program been effective in selecting grantees within categories best able to provide practice-relevant findings?

Have grantees been encouraged and supported to maximize project success?

In understanding project effectiveness, have teachers and learners had a voice?

Is the program sensitive to and implementing projects that result in disseminatable findings?

In sum, the questions should cover not only what has been accomplished within the program but for whom those accomplishments apply and under what conditions. If the answers to the questions are derived through a collaborative process engaging representatives of the various audiences in a consensus-building process, the results are more likely to be applicable to the educational system and not fragmentally to one or another part of the system.

One Strategy for Collaborative Evaluation—A Brief Description of Cluster Evaluation

What follows is a description of an evaluation method that engages a group—or cluster—of projects in common evaluation efforts. Using this method, the authors have been able to accomplish the purposes discussed earlier for NSF evaluation. Cluster evaluation provides a complex, rich data set, derived to a large extent from the involvement of stakeholders in the formation of the evaluation itself, that provides information for determining program impact, as well as improving programs. The process of the cluster also enables and prepares project directors to improve their own evaluation skills and allows them to be better consumers of evaluation data. The authors believe the cluster evaluation model has widespread application in the NSF arena.

The generic method of cluster evaluation was described and named by the W.K. Kellogg Foundation and is used in their various funding initiatives. Implementation, however, varies from cluster to cluster. The specific cluster evaluation method developed and used by the authors with two groups of 12 science education projects is summarized below.

Organizing the Cluster

The specific organization of the cluster is affected by several factors, including the number of projects funded, geographic location of projects, nature of topical area, targeted populations, and degree of similarity of the project implementations. Availability and level of experience of the cluster evaluators also affects the process.

Selection of cluster evaluators is initiated by the funder, and basic organization, time frame, role of funder program staff, evaluators, and project staff, and implementation procedures for the cluster evaluation are negotiated.

Projects selected for inclusion in the cluster are usually determined by the funder. Completion of selection of projects varies, with some selected prior to the initiation of the evaluation and others selected several months into the process. Based on the authors' experiences, selection prior to initiation of the cluster evaluation results in a more effective evaluation.

The number of projects in a cluster can vary, depending on the factors described above. The basic purpose and expected results of the cluster evaluation should be carefully considered, along with available financial and other resources. Clusters of not more than 25 are optimal for conducting an intensive collaborative cluster evaluation as described in this paper.

Regular networking conferences are organized by cluster evaluators and program staff, with funding included in cluster evaluator budgets or a separate budget. Additionally, resources must be made available to funder-program staff to participate in conferences and technical assistance.

A retrospective cluster evaluation of completed projects is also possible, but would necessitate assembling directors from completed projects. The purpose and results of a retrospective cluster evaluation would be different from one with a formative emphasis.

NSF research-oriented projects, such as those in the RTL program, could be easily placed in clusters based on a set of factors, from topic to implementation strategy, and determined by specified evaluation purposes. A retrospective cluster could be determined by regional or other representative sampling techniques.

> *"... cluster evaluation is a complex process with diverse components requiring a variety of skills and resources ..."*

Cluster Evaluation Team

Because cluster evaluation is a complex process with diverse components requiring a variety of skills and resources, a team of evaluators should be enlisted. It should include people with evaluation expertise, research skills, human relations skills (including writing skills), and appropriate content-area knowledge. Additionally, adequate support staff must be available to attend to details of networking conferences, data collection/compilation, communications, etc. Although not all team members necessarily have to devote full time to the effort, sufficient professional staff time must be available to coordinate and carry out the many evaluative tasks.

In the case of the science education cluster evaluations conducted by the authors, the cluster evaluation team is made up of two principal investigators, one with a strong background in research and evaluation, the other with extensive experience in science education. Additionally, doctoral students and staff bring research, evaluation, organizational, and communication skills to the team. Keeping current in the content area is necessary if evaluators are to provide useful information to improve programs and to judge outcome accomplishment.

Additionally, external content area and evaluation specialists should be enlisted to periodically review the cluster evaluation.

Setting Expectations

It is important to set expectations for the cluster evaluation up front not only for funders and cluster evaluators, but also for project directors and their staff. Although some projects may have a proposed evaluation plan, including an internal evaluator to implement it, most will need assistance with both internal and cluster evaluation activities. Expectations for funded projects must include full participation in all cluster evaluation activities, including networking conferences, data collection and analysis, and reporting and dissemination. Funders must make these expectations clear and provide adequate resources to facilitate full participation.

Through RFPs or in award letters, NSF staff would make expectations clear for full participation in the cluster evaluation. Additional communications would introduce cluster evaluators and provide instructions for collaboration.

Negotiated Common Cluster Outcomes

Usually at the first networking conference following selection of projects for the cluster, initial common cluster outcomes are determined collaboratively. Using important evaluation questions developed by project and funder program staff for specific projects and questions developed by cluster evaluators and program staff for the overall cluster, a comprehensive list of outcomes is devised. From this list, a set of common cluster-level outcomes is developed by consensus of the project directors and evaluators, funder program staff, and cluster evaluators. In one science education cluster, 19 cluster outcomes, held in common by two or more projects, were created addressing issues related to students, teachers, curriculum, collaboration, and continuation/ dissemination.

As projects evolve and the cluster evaluation develops, modifications are made to the common cluster outcomes as appropriate, such as adding outcomes or modifying existing ones to better reflect actual intended outcomes. This set of outcomes provides a partial framework for the evaluation of the cluster of projects, and "represents to the projects the intended impact of the cluster" (Barley, 1991). Individual project-level evaluations may also be conducted by projects in the context of the cluster evaluation, depending on requirements of the funder.

> *"As projects evolve and the cluster evaluation develops, modifications are made to the common cluster outcomes as appropriate..."*

For use at NSF, staff, along with cluster evaluators, would develop a set of important questions for the overall evaluation of, for example, a cluster of RTL projects. Some questions will be pertinent to the overall RTL program and others specific to the particular cluster of RTL projects. Individual project staff develop important questions pertinent to their own projects. Collaboratively, a set of common cluster outcomes is then established through negotiation.

Collaborative Data Collection

Both qualitative and quantitative data come from a variety of sources and are in a variety of forms. Individual projects collect data directly from the participants through questionnaires, interviews, observations, journals, standardized tests, recordkeeping, and common cluster instruments (same instruments used across projects to collect consistent data). Some data are reported in annual reports; other data are sent directly to cluster evaluators. Cluster evaluators collect data from cross-project participant surveys, project staff interviews, documents, participant interviews, and site visits and observations. Also collected is specific information on the strategies and activities each project uses to accomplish the cluster outcomes, as well as contextual information pertinent to cluster outcomes.

> *"When expectations for data collection are clear early in the process, ... better data are the result."*

Several factors affect the quality and quantity of data, including commitment of the various stakeholders to the process, financial resources, and data collection design. When expectations for data collection are clear early in the process, and cluster evaluators facilitate the process through technical assistance and instrument development, better data are the result.

It would be important for projects within an NSF cluster to collect data pertinent to individual project and cluster outcomes, as well as contextual factors and implementation strategies. With technical assistance from cluster evaluators, project directors and their staff will be in the best position to collect pertinent data for individual project and cluster use. Cluster evaluators would also conduct across-project data collection efforts.

Regular Networking Conferences

Direct networking among all project directors, project staff and evaluators, cluster evaluators, funder program staff, and guests at annual or semi-annual networking conferences is an important component of cluster evaluation. The purposes of these conferences will vary somewhat depending on the purpose of the evaluation, topical focus of the cluster, and frequency of the meetings. All networking conferences should include sessions (1) to conduct strategic planning for, exchange ideas about, provide direction to, discuss issues and problems emerging from, and review and analyze data and findings of the cluster evaluation; (2) share lessons learned with other projects; and (3) visit project sites. For a science/mathematics education focused cluster, for example, purposes should also include learning about current and developing issues in science education and science curriculum, instruction, and assessment topics directly pertinent to projects; and formally and informally sharing science education curriculum materials and instructional strategies. Networking is at the heart of a constructivist approach, since it provides a forum for direct engagement of major stakeholders in the cluster evaluation process.

Networking conferences are organized collaboratively between cluster evaluators, program staff, and project directors. Specific travel, overnight accommodation, meal, and meeting arrangements can be part of the cluster evaluator's responsibility and funded accordingly, or the funder can arrange or contract for networking conferences. The number and duration of conferences are related to the purpose of the cluster evaluation and/or available financial resources.

For a cluster evaluation of NSF projects, program staff would be actively involved with cluster evaluators in planning and implementing the conferences. Project directors, individually and in committees, provide feedback and can help make arrangements for the gatherings.

Data Analysis and Working Hypotheses

A method used in one of the authors' science education clusters to review and analyze the diverse outcome-related data is the use of "working hypotheses," a term first coined by Cronbach (1975), describing tentative hypothesizing statements "that give proper weight to local contextual conditions," but facilitate transferability across varying contextual situations. The degree of transferability depends on the similarity between contexts—the "fittingness" or "degree of concurrence between sending and receiving contexts" (Lincoln and Guba, 1985). Core and auxiliary working hypotheses, based on common cluster outcomes, address commonalties and differences in project-level implementation strategies (Barley, 1991). Working hypotheses are reviewed and modified at networking meetings. Tentative findings are developed by the evaluators and presented to the cluster members for further review. Project staff have an opportunity to offer suggestions for modifications based on additional data and findings from individual projects and make recommendations for additional relevant data collection.

Other analysis methods for mixed data can be used, but should involve project directors and staff at appropriate points in the process.

Cooperative Derivation and Dissemination of Results

Dissemination of findings and sharing of lessons learned occurs between individual projects in the cluster, from individual projects to other pertinent programs (for example, science/mathematics education programs for an NSF cluster), among projects at networking conferences, and at local, state, and national gatherings of educators, evaluators, and others. Networking conference sessions are also devoted to planning common dissemination activities, such as development of printed materials, videos, conferences, consulting services, etc.

This will be an important aspect of cluster evaluation for NSF programs, since project directors and staff must be actively involved in deriving and disseminating results, not only of the evaluation, but of project research findings. Evaluation findings should help NSF program staff, in collaboration with cluster evaluators and project directors, determine future funding and research efforts. Networking within a cluster and between clusters would also facilitate interactions among a large group of researchers and NSF staff, leading to more informed coordination of NSF-funded research activities and their relationship to overall education reform efforts.

"... cluster evaluation engages stakeholders in the evaluation process."

Recommendations and Conclusions

Cluster evaluation as briefly described in this paper is an innovative and effective method that can be appropriately adapted to help meet the needs of the National Science Foundation as it seeks to develop an evaluation framework that will identify the footprints left behind by its programming efforts. Although cluster evaluation can be used retrospectively, it is particularly appropriate when used with groups of projects initiating and conducting their programs, thus identifying footprints throughout the course of the projects.

As a formative/summative combination approach (as described in this paper), cluster evaluation engages stakeholders in the evaluation process. It provides feedback to projects as they implement their programs, and, thus, helps them improve. Cluster evaluation also measures the overall impact of the group of projects and addresses contextual factors and implementation strategies.

Using it retrospectively, cluster evaluation provides a framework for addressing important evaluation questions related to outcomes, context, and implementation. It is suggested that an evaluation "panel," representative of a broad cross-section of NSF stakeholders, project directors, program staff, evaluators, teachers, and learners, be established for particular NSF program areas or portions of a program area (i.e., projects with similar missions). Operating collaboratively and on an ongoing basis, their purpose would be to construct and adjust the evaluation design out of their shared concerns, values, and directions for the program. They would jointly establish the evaluation questions, determine the specific design, collect common data, and develop analyses appropriate to the real world of educational practice.

References

Barley, Z.A. 1991. Strengthening community-based project evaluations and deriving cross-project findings. Paper presented at the meeting of the American Evaluation Association, October 1991, Chicago.

Bellavita, C., Wholey, J.S., and Abramson, M.A. 1986. Performance-oriented evaluation: Prospects for the future. In *Performance and Credibility: Developing Excellence in Public and Non-Profit Organizations*, J.S. Wholey, M.A. Abramson, and C. Bellavita, (eds.). Lexington, MA: Lexington.

Cronbach, L.J. 1975. Beyond the two disciplines of scientific psychology. *American Psychology* 30(2): 116-27.

Cronbach, L.J., Ambron, S.R., Dornbusch, S.M., Hess, R.D., Hornik, R.C., Phillips, D.C., Walker, D.F., and Weiner, S.S. 1980. *Toward reform of program evaluation.* San Francisco: Jossey Bass.

Donmoyer, R. 1991. Postpositivist evaluation: Give me a for instance. *Educational Administration Quarterly* 27(3): 265-96.

Guba, E.G., and Lincoln, Y. S. 1989. *Fourth generation evaluation.* Newbury Park, CA: Sage.

Lincoln, Y.S., and Guba, E.G. 1985. *Naturalistic inquiry.* Newbury Park, CA: Sage.

Wholey, J.S. 1983. *Evaluation and effective public management.* Boston: Little, Brown.

Wholey, J.S., and White, B.F. 1973. Evaluation's impact on Title I elementary and secondary education program management. *Evaluation* 1: 73-76.

The Virtual Reality Of Systemic Effects Of NSF Programming On Education: Its Profession, Practice, Research, And Institutions

Robert E. Stake
University of Illinois

It is both healthy curiosity and political necessity to wonder how research and development in science education is affecting not only the teaching and learning of science but also the greater educational and social system. In this paper, I review concerns about program effectiveness and accountability, and comment on the capabilities of program evaluation methods and people to trace systemic effects. Before identifying potential contributions from qualitative methodology, I outline its common characteristics. Claiming an interpretive commitment to be qualitative research's characteristic most applicable here, I suggest creation of, for each major program of the directorate, a semi-independent evaluation council for long-term interpretive study of the systemic influence of NSF educational research and development on various fields of action.

Seeking New Strategies for Program Evaluation

Thirty years of experience with the evaluation of Federal programs has persuaded many members of the American Evaluation Association that "there are no easy answers." At each year's annual meeting, there are restatements of the perplexity and renewed attention to political and cultural contexts. The foundation for future strategic thinking should not ignore the presidential addresses, the 96 theses of Lee Cronbach and colleagues (1980), and the 31 "hard-won lessons" identified by Michael Scriven (1993). Applying some of the experiential wisdom expressed in those resources to the present task, I begin with the following 17 caveats.

Evaluation Strategies: Caveats

1. Providing indicators of program impact is a task fraught with political and promotional pressure, resulting in overly "favorable" evaluations (Scriven 1991), resulting in evaluation schemes that exceed the technical capacities of evaluators. Realistic review of evaluation strategies is uncommon. Over-promising becomes routine. Organizational structures should be developed to require more realistic strategies for evaluating NSF programs.

2. Efforts to measure program merit and effect face complex political environments that reward:

   a. Delaying action (evaluation seldom can happen fast enough to support or counter impressions and experiences of the program itself);

   b. Disguise of advocacies (by reifying certain criteria of success and obscuring others, groups oriented to the reified criteria are supported); and

   c. A facade of accountability (the act of commissioning an evaluation makes it appear that the commissioning agency is acting responsibly).

New strategies need to be directed as much at disengaging evaluation from the advocacies of science and mathematics education as at finding new representations of effect.

"New strategies need to be directed as much at disengaging evaluation from the advocacies of science and mathematics education as at finding new representations of effect."

3. While group efforts to examine strategies for program evaluation should be encouraged, strategies are not necessarily strengthened by group consensus. Strength is also to be found in a diversity of ideas. It may be more important to add strategic options, some unpopular, to the armamentarium than to find a grand strategy that has few opponents.

4. Uniform strategies across programs is not an important end. Dissimilarity within and between programs requires nonuniform evaluation methods. If methods are too dissimilar, understanding of program effects will be low. With strategies too similar, unique contributions of individual programs will be understated (Cronbach, et al. 1980).

5. One strategy recognized almost universally is that multiple measures of important constructs are highly desirable. Conducting multiple studies is one way of getting multiple measures, some of which will help validate the constructs and others which will help illustrate the different interpretations given a construct in different settings.

6. Evaluation data can be newly generated by research or can be gathered from people who already are interpreting what is happening.

7. Most government-sponsored evaluations are designed in instrumentalist fashion; that is, the program is presented as an agent effecting some change in operations and productivity with certain benefit to a clientele. In the eyes of many advocates and clients, however, program quality is seen as the quality of services provided, as intrinsic quality rather than product quality. The social sciences are a reservoir of instrumentalist views; the humanities are a reservoir of intrinsic views. A review of evaluation strategies should consider both (Guba and Lincoln 1981).

8. Whether or not programs should be evaluated formally is a political and administrative matter more than a developmental and epistemological matter. It is common knowledge that formal evaluation studies have usually not provided critical input to government decision making about continuation or change in programs.

9. Evaluation occurs both formally and informally. Those closest to the scene tend to be more satisfied with informal than formal evaluation. People at a distance, especially those dubious about the program, tend to prefer formal and independent evaluation.

10. Most programs supported by the National Science Foundation are complex. Instruction and other discourse affected by NSF programs are simultaneously being affected by many other influences. Attribution of effect to NSF programs is problematic, at best.

11. The more distant an intended effect is from program activity, the more difficult the attribution. Distance can be a matter of time, place, personal interaction, content, or conceptualization.

12. The pre-announced metaphor of "footprints" as an indicator of effects of a program's passing should be given no more than a moment's thought. That metaphor raises the image of pristine sur-

faces, such as newly waxed floor or fresh sand at the beach, and the fitting of a slipper to Cinderella-like program agents. Real surfaces are scuffed, trammelled, and exposed to countless footfalls, and real programs rarely leave distinguishing marks. But the major flaw in the metaphor is its romantic image of an indicator that requires little human interpretation.

13. Education and human beings are extremely complex. We seldom can measure effects of educational research and development directly. Validity of measurement tends to diminish, the more indirect the indicator. For a nation, a school, and sometimes even a child, our indicators of program effect are quite indirect. Many are of low validity. Indicating the systemic effects of NSF programming on research, training, professional communication, and popular discourse directs attention to quite indirect outcomes.

14. We have indicators of high validity and those of low (Shavelson, et al. 1987; Guiton and Burstein, 1993). Misleading evaluations result from interpreting indicators beyond the limits of their validity. For example.

These indicators:	**are a good indicator of:**	**but a poor indicator of:**
need statements	what people would like	what is actually needed
standardized test scores	student ability	actual student achievement
grade point averages	compliance in instruction	ability to use own knowledge
courses taken in Education	teacher formal qualification	teaching quality
monetary costs	money spent	the social costs
followup ratings	participant satisfaction	program effectiveness

15. Indicators have a reactive effect. To get better test scores or other marks, schooling is redirected to better affect the indicator variable, sometimes at the expense of the real targets of education. Both insiders and outsiders increasingly substitute the indicator variable as the definition of education. Were we to create valid indicators of systemic effects, advocates and adversaries would probably find ways to subvert them.

16. Essentially, in evaluation studies, we are not aiming as much to identify program effects as to identify the value of the effects (Scriven, 1991). Value of effect requires consideration of costs. (In education, worth and costs are seldom measured in dollars.) At least as hard to measure as effects, values and cost measurements are seldom included in an evaluation design. Strong measurement designs often presume that values and costs will be apparent without measuring. Sometimes the best strategy will be to obtain summary judgments from people who themselves have been exposed to all three.

17. Increased attention is being given to the design of indicators of provision of educational opportunity. School delivery standards (Porter 1993) would change evaluation strategy to concentrate more on the measuring of process and less on the measuring of product. A strategy emphasizing systemic effects runs counter to emphasis on provision of opportunity.

"*For most people, the evaluation of Federal programs raises the expectation that something will be measured to which a value can be attached.*"

I open my paper with these 17 caveats intending to help anchor discussions of evaluative strategy in practical experience. I think it is possible to increase NSF sensitivity to the effects its programs are having, but precise, validated, and immediate indicators are some of the illusory "easy answers." How NSF sensitivity and program advocacy may be enhanced by nontraditional evaluation strategies requires a careful look at what is expected of program evaluation.

What Is Being Asked of Evaluation

Essentially, evaluation is the determination of merit and shortcoming (Scriven, 1967). Program evaluation is commonly taken to be "systematic examination of events occurring in and consequent on a contemporary program ... to assist in improving this program and other programs having the same general purpose" (Cronbach, et al., 1980). For most people, the evaluation of Federal programs raises the expectation that something will be measured to which a value can be attached. (In this paper, I am not speaking of project or proposal evaluation but the evaluation of large NSF programs, especially their effects on the educational R & D enterprise and on education generally.)

A Contrived Rationality- Program evaluation, like the social sciences, is in the business of making rational what is empirical. Our principal knowledge of life is empirical. Although indirect and sporadic, much of our knowledge of the work of government is empirical. We try to rationalize what we experience. Government programs change, society changes, people change, all calling for changes in our rationalizations.

Evaluation specialists get contracts to discern a program's measurable relationships, particularly cause and effect relationships. And most evaluators confidently try—operating under the notion that if change has occurred, a cause can be discerned. If subsequent conditions seem to connect back to the program more than to anything else, then it may be said that the program caused the effects. Proof of such a relationship is far beyond reach. Certainly, in program evaluation, if not everywhere, cause and effect is a constructed reality—sometimes a contrived reality.

The context of government programs is political. Information needs are unlike contests common to researchers (Chelimsky, 1991). Problems are real and taken seriously, but expediency and irrationality are common. Almost every government official is tuned to the morning news (Barnouw, 1970). Bureaucracies strive for rationality; failing that, for the appearance of rationality. They are beset by news media not only for news but for stories. The media are presumptuous about rationality. They equate rationality with responsibility. They imply rationality to be the responsibility of officials, whose information systems are expected to tell what is causing what.

Reporter orientation to causality is particularly aroused by a calamity such as the immolation of the Branch Davidian cult in Waco, Texas. Did the

FBI provoke a mass suicide? Did the President really take full responsibility? Looking back on the Waco calamity, columnist Michael Kelly of the Washington Post discerned the discrepancy between public and media stances, noting little interest within the public in finding someone to blame (April 1993). Kelly used words of Robert Coles, which described the media's "... arrogant faith in rationalism ... , all of them paying homage to the great delusion of our times, that social scientists will deliver us from irrational madness and the random hand of fate." Blame makes a good story. Under media expectations, it behooves evaluators to identify blame for program shortcomings.

Deliverance also makes a good story. Within professional education at present, much attention is paid the Curriculum and Evaluation Standards for School Mathematics, published by the National Council of Teachers of Mathematics (NCTM) in 1989. Does problem-solving get graduates ready for the work place? Is NCTM now leading the school reform movement? Some believe evaluators should be trying to measure such effects. How should they evaluate the effects of NCTM Standards? Perhaps by looking into other teaching areas (Ball, 1992). Specialists in language arts promoting a "whole language" approach occasionally mention the NCTM Standards. Specialists in distance education trying to develop simulations far from campus occasionally mention the NCTM Standards. Is their work influenced by the Standards? Possibly, but not on the basis of how frequent is the mention or how congenial the innovation. Workers in other fields see that the legitimacy of the Standards might rub off on their efforts, so they cite them. Citation does not mean they have been influenced by the mathematics teachers.

Now that we have thought about it, there may be a phenomenon we can call the NCTM effect on school improvement. And an evaluator might be able to estimate how much the work of mathematics teachers has influenced other innovatory efforts. Could we call the estimate an indicator? Could we validate the estimate? Indicator validation is not going to happen. The estimate itself may be useful, not only for promotional purposes, but in the rumination and discourse of program management. But estimates are not facts. Indices such as "the NCTM effect" or "readiness for the work place," just like the now vernacular "employment rate" and "Dow Jones average," however useful, are fictions, beyond constructed realities, a form of that new whiz bang, "virtual reality." More on that in a moment.

The real work of educators is not "to look good," nor is it "to catch up with the Japanese," nor is it even (in my view) "to cause the child to be different," but to provide opportunity and pressure for children to follow preferred paths to becoming educated. It is the natural state of the child to be affected by teachers and tenuously by distant research programs. How much the separate layers of the system can take credit for good effects—or bad, for that matter—is beyond the understanding of everyone, including evaluation specialists. Whatever the appetite for indicators, whatever the demand for program accountability, however useful measurement of effect might be, the state of the art is such that indicators of systemic effect are not available. And it is irresponsible for officials to use unvalidated indicators of effect as if they had been validated. And it is an act of deception for evaluators to provide such indicators.

What state-of-the-art evaluators can do is to see if programs are drawing upon the best of human understandings, organizing programs in felicitous ways, recognizing and coping with problems, maintaining a dignity of human relations. It is not wrong to be curious about outcomes, but it is wrong to join in the deceit that governments cause education, and in the self-deceit that evaluators reliably measure and attribute effects. It is wrong to portray a rationality that does not exist.

"There is no single wellspring of qualitative research from which to draw methods for evaluating NSF strategies."

It is also wrong to base evaluation strategy on what ought to be rather than on what is. Formal evaluation expectations are based largely on specialist services presently available. They do evolve, and can be seen to be increasing their use of qualitative field work, particularly with case studies and ethnographic interpretations. How NSF sensitivity and program advocacy may be enhanced by nonresidential evaluation strategies requires more than a passing knowledge of qualitative research methods. Drawing upon the *Handbook of Qualitative Research,* (Denzin and Lincoln, 1994), the following section is my distillation of that emerging methodology—disciplined qualitative inquiry.

The Nature of Qualitative Research

There is no single wellspring of qualitative research from which to draw methods for evaluating NSF strategies. Practices vary in different fields. The distinction between quantitative and qualitative methods is a matter of emphasis more than a matter of boundary. In each ethnographic or naturalistic or phenomenological or hermeneutic or holistic study, i.e., in each qualitative study, enumeration and recognition of differences in amount have a place. And in each statistical survey and controlled experiment, in each quantitative study, natural-language description and researcher interpretation are expected. Perhaps the most important differences in emphasis are threefold:

a. Distinction between knowledge discovered and knowledge constructed;

b. Distinction between aiming for explanation and aiming for understanding; and

c. Distinction between personal and impersonal roles of the researcher.

Constructed Knowledge and Virtual Representations- The children of today are manyfold the linguists their parents were as children. Their exposure to images has grown a hundredfold. Their access to keyboards and software gives them vast new ranges of expression. Literary empowerment has been enormous for evaluators as well. We can say so much more, represent it in so many more ways, prepare handsome camera-ready copy ourselves.

As the electronic field has exploded in both sophistication and public access, the art of representation has exploded too. Readers can be immersed in the description, drawn into the most elaborate of vicarious experiences (Spiro, et al. 1987). Following Aldous Huxley's *Brave New World,* broadcast advertising (Barnouw, 1970), and, more recently, computer artist Myron Krueger's Artificial Realities (1983), we are passing into a period of interactive stimulation that extends personal experience far beyond the movies and charismatic teaching. Among its champions, it is called, "virtual reality" (Woolley, 1992), making possible a sensory reality beyond ordinary experience, such as playing tennis on a low gravity court. Radio talk shows have been titillating the public with ideas about simulating pleasure. A few "virtual reality" venues are more sober, more intellectual. A number of our colleagues in artificial intelligence research have designed extra-responsive environments for simulation of problem situations to enhance learning (Psotka, 1993). But this medium is one of grand deception. As Lewis Carroll explained, "For the snark was a boojum, you see."

What I said two paragraphs back about empowerment of children and evaluators is merely an assertion, another virtual reality, but one I expect will sit comfortably with most readers. If that claim is not true, it is virtually true. It is an untruth most people will accept as true. Increasingly we realize our dependence on virtual truths. We pause in our own metamorphosis. As we increase our ability to represent the world, we have greater difficulty in remembering what the world actually has been, and increasing doubt we ever knew what it might be. Some virtual realities we settle for, some we aspire to, such as those we call science and art. We cannot even imagine a world without these virtual realities, these constructs, these indicators. Our problem is one of believing them too much, losing the appetite for validation.

Multimedia shows and role playing repeatedly have shown us that simulation creates a reality of its own. When simulation is effective, that which was simulated can become secondary to the simulation. Shakespeare and McLuhan agreed, "The show's the thing." Virtual sunsets outdo the real in so many ways. The classical questions reappear: "What is reality?" "Is there substance behind appearance?" Children and researchers create new knowledge. And in telling others what they have learned, even as they remember, they simulate that knowledge. New knowledge and simulated knowledge are different (Stake and Trumbull, 1982), propositional and tacit knowledge are different (Polanyi, 1969), but I often find them difficult to tell apart.

In our personal lives, some symbols, narratives, and indices stand for the real thing, more stand for other symbols, narratives, and indices. We remember, sometimes remembering memories rather than the original experience. We create within our minds a world of representations. We do this from our earliest ages, seeking to make sense of puzzling environments, repeating experiences, refining our indicators—but all too seldom do we go out of our way to validate them.

"As we increase our ability to represent the world, we have greater difficulty in remembering what the world actually has been, and increasing doubt we ever knew what it might be."

In our societal and institutional lives, we of course need symbols, narratives, and indices. We do not know how to survive without them. We are jolted by the realization that indices are created for other purposes than representation: as dreams and icons, as subterfuge, as enhancements and caricatures, as provocations and supplications. Secretary of Education Terrell Bell created his famous Wallchart of SAT scores ostensibly to represent the quality of secondary education in the 50 states. He knew the data were greatly misleading, but posted them to provoke researchers into creating a valid comparison (Bell, personal communication). Indices exist for advocacy as much as for information. New indices are seldom validated over a developmental period before being offered for public or specialist interpretation. It is part of our evolving language, part of our evolving knowledge base, to have grand indices, but it is part of our carelessness to take them to mean what they seem to mean.

A preponderance of qualitative methodologists are constructivists, professing belief that knowledge is the invention of inquiring minds, not their discovery (Schwandt, 1994). Knowledge is made, not found. Qualitative study of teaching and learning correspondingly emphasizes the construction of ideas by children rather than the acquisition of ideas. This is not just a preferred set of learnings or preferred pedagogy, but an epistemological definition. Each person constructs knowledge, most not recognizably unique, but individually created. We have common knowledge not so much because there are pre-existing facts, truths, for us to discover, but because learning, like dressing and driving, is a social process. We have strong tendencies to conform. We modify our actions to fit the actions of those we respect. And we create knowledge that appears to be very similar to that of the people around us.

The important thing to the qualitative researcher is that it is helpful to consider much learning, much "reality," as human construction. It is necessary sometimes to be reminded that the indices, the virtuals, we use to monitor our lives are contrivances regularly in need of calibration.

Experiential Understanding- A distinction among aims, an epistemological distinction, fundamentally separates qualitative and quantitative inquiry. The distinction is not that between quantitative and qualitative data. The distinction is in intent, between inquiry for making explanations versus inquiry for promoting understanding. It has been nicely stated by philosopher George Henrik von Wright in his book, *Explanation and Understanding* (1971). Von Wright recognized that understanding is personally constructed. He acknowledged that explanations are intended to promote understanding and understanding is often expressed in terms of explanation—but epistemologically, the two are quite different. Von Wright emphasized the difference between generative explanation and experiential understanding.

It is a distinction seen in preferences for process versus product evaluation. Products are the manifestation of generative processes. Choosing product evaluation is problematic for us because the causes of systemic effects are not necessarily the processes we assume, allude to, or experience. Given such uncertainties, the qualitative evaluator gives greater attention to process as experienced (Guba and Lincoln, 1982), with the reader expected to share in the interpretation. For the educator, the distinction parallels the difference between preparing to teach didactically and preparing experiential opportunities for learners. Shall researchers tell a reader what they have learned, or shall they arrange a situation optimally suited to reader learning? Qualitative evaluation designs generally aim to have evaluators make descriptions and situational interpretations of phenomena, which they offer colleagues, students, and others for modifying their own understandings of program merit (that is, for "naturalistic generalization," as Deborah Trumbull and I called it in 1982). Quantitative evaluation designs generally aim to advance abstract comprehensions of the evaluators who, in turn, present these explanations to their colleagues, students, and diverse audiences.

Qualitative descriptions are expected to be recognizable by readers, yet no description captures veridically the phenomena described. Jorge Luis Borges spoke of this elusive character of language in *A Yellow Rose*:

> ...Then the revelation occurred: Marino saw the rose as Adam might have seen it in Paradise, and he thought that the rose was to be found in its own eternity and not in his words; and that we may mention or allude to a thing, but not express it...

Borges recognized the inescapable artificiality of description.

Quantitative research methods have grown out of search for grand theory. To establish generalizations that hold over diverse situations, most social science-oriented researchers make observations in diverse situations. They try to eliminate the merely situational, letting contextual effects "balance out." They try to nullify context in order to find salient and pervasive explanatory relationships. Qualitative evaluators treat the uniqueness of individual contexts as important to understanding.

Most program evaluation work has been dominated by science's search for grand explanation. Employment of formal measurement and statistical analysis, i.e., quantification, has occurred in order to permit aggregation of a large number of dissimilar cases, thus to position the researcher to make formal generalizations about the program. The appropriateness of scientific explanation for program evaluation has been questioned by Scriven (1978) and Cronbach (1980, et al.) on the grounds of the particularity of the evaluand, its situationality, and its political context. Both of them have emphasized the evaluator's responsibility for authoring program-specific descriptions and interpretations. Practicing evaluators draw upon both quantitative and qualitative methods, choosing one or the other to provide scientific explanation or experiential understanding.

Emphasis on Interpretation- Qualitative evaluation specialists such as Elliot Eisner (1979) and Egon Guba and Yvonna Lincoln (1981) have urged reliance on direct interpretation of events more than on interpreted measurement of attributes. All research designs feature interpretation—but, with standard quantitative designs, there is effort to constrain interpretation during that period extending from design of the study to analysis of the data. Standard qualitative designs call for the persons most responsible for interpretations to be in the field during that period, responding to the situation (Stake, 1975), making observations and interpretations simultaneously.

"Practicing evaluators draw upon both quantitative and qualitative methods, choosing one or the other to provide scientific explanation or experiential understanding."

The difference is epitomized by two kinds of research questions. In quantitative studies, the research question typically embodies a relationship among a small number of variables, e.g., "Is there an enduring correlation between applicability of technological support and teacher qualification over a variety of situations?" Efforts are made to operationally bound the inquiry, to define the variables, and to minimize the importance of interpretation until data are analyzed. At the very beginning, it is important to anticipate how relationships between variables would reduce weakness in explanation and, at closing, it is important for the researchers to modify their generalizations about the variables. In between times, it is important not to let interpretation change the course of the evaluation study (Stake,1994).

In qualitative studies, the research question typically orients to cases or phenomena, seeking patterns of unanticipated as well as expected relationship. For example, "What will happen to collegial relationships among teachers working with this program if all are obligated to emphasize a problem-solving pedagogy?" Or if the project had been implemented sometime in the past, "What happened?" Dependent variables are seldom operationally defined, situational conditions may not be known in advance, even the independent variables are expected to develop in unexpected ways. It is important to have the interpretive powers of the research team in immediate touch with developing events and ongoing revelations, partly to redirect observations and to pursue emerging issues. The allocation of resources is different. Reliance on carefully developed instruments and redundancy of observations typical in a quantitative study give way to placement of the most skilled researchers directly in contact with the phenomena and making much more subjective claims as to the meanings of data.

"Thick description, alternative interpretations, 'multiple realities,' and 'naturalistic generalization' are not only common; often they are aims for these nontraditional research methods."

In his fine summary of the nature of qualitative study, Frederick Erickson (1986) claimed that the primary characteristic of qualitative research is interpretation. He said that findings are not just "findings" but "assertions." Qualitative study is not alone in personalizing interpretation. Speaking of all social science, Henry Aaron (1978, 156) claimed:

> Outsiders may be lulled into thinking that issues are being debated with scholarly impartiality, when in fact more basic passions are parading before the reader clad in the jargon of academic debate.

Qualitative methods invite personal reflection. With intense interaction of researcher and actors in the field, with a constructivist orientation to knowledge, with sensitivity to participant intentionality and sense of self, however descriptive the report, the qualitative researcher expects to express personal views.

Erickson drew attention to the ethnographers' traditional emphasis on emic issues, those concerns and values recognized in the behavior and language of the people being studied. Geertz (1973) called it: "thick description." And often the aim is not veridical representation so much as stimulation of further reflection, optimizing readers' opportunity to learn. Stake and Trumbull (1982) called it "naturalistic generalization," a concern for assisting the reader's further understandings. It draws from history, philosophy, and literature, sometimes paralleling the artist's work. Claude Debussy, on composing La Mer, not at sea, but in his Paris studio, said:

> I have my memories and they are better than the seascapes themselves whose beauty often deadens thought. My listeners have their own store of memories for me to dredge up.

The function of research is not always to map the world but to sophisticate the beholding of it.

Thick description, alternative interpretations, "multiple realities," and "naturalistic generalization" are not only common; often they are aims for these nontraditional research methods. Such pursuit of complex meaning cannot be just designed in or caught retrospectively (Denzin and Lincoln, 1994). It seems to require continuous attention, an attention seldom sustained when the dominant instruments of data gathering are objectively interpretable checklists or survey items. An ongoing interpretive role of the researcher is prominent in the work of qualitative research.

Other Characteristics of Qualitative Research- In addition to its orientation away from cause-and-effect explanation and toward personal interpretation, qualitative inquiry is distinguished by its emphasis on holistic treatment of phenomena (Schwandt, 1994). I have remarked already on the epistemology of qualitative researchers as existential (as opposed to causal or generative) and constructivist. These two views are correlated with an expectation that phenomena are intricately related to many coincidental actions and that understanding them requires a wide sweep of contexts: temporal and spatial, historical, political, economic, cultural, social, personal.

Thus the case, the activity, the event, is seen as critically unique as well as common. Understanding it requires an understanding of other cases, activities, and events. Uniqueness is recognized not primarily by comparing cases on a number of variables—there may be few ways in which this immediate case strays from the norm—but the collection of features, the sequence of happenings, is seen by people close at hand to be in many ways unprecedented and important; that is, a critical uniqueness. Readers are drawn easily to a sense of uniqueness as they read narratives, vignettes, experiential accounts (van Maanan, 1988). The uniquenesses are expected to be critical to the understanding of the particular case.

For all their intrusion into habitats and personal affairs, qualitative researchers are non-interventionists. In the field, they try not to draw attention to themselves or their work. Other than positioning themselves, they avoid creating situations to test their hypotheses. They try to observe the ordinary and they try to observe it long enough to comprehend what, for this case, ordinary means. For them, naturalistic observation has been the primary medium of acquaintance. When they cannot see for themselves, they ask others who have seen. When formal records have been kept, they scrutinize the documents. But they favor a personal capture of the experience, so they can interpret it, recognize its contexts, puzzle the many meanings, while still there, and pass along an experiential, naturalistic account to allow readers to participate in some of the same reflection.

Recognition of Risks- Qualitative study has everything wrong with it that its detractors claim. It is subjective. The contributions toward an improved and disciplined science are slow and tendentious. New questions are more frequent than answers. The results pay off too little in the advancement of social practice. The ethical risks are substantial. And the costs are high.

The effort to promote a subjective research paradigm is deliberate. Subjectivity is not seen as a failing to be eliminated but as an essential element of understanding. Still, personal understanding frequently is misunderstanding, by actors, by the researchers, and by readers. The misunderstanding may occur because of the intellectual shortcomings of the interpreter or because of weakness in protocol which fails to purge misinterpretation. Qualitative researchers have a respectable concern for validation of observations, they have routines for "triangulation" (Denzin, 1989) that can approximate in purpose those in the quantitative fields, but they do not have the protocols that put subjective misunderstandings to a stiff enough test.

Many phenomena studied take long to happen and evolve along the way. Often we need a long time to come to understand what is going on. The work is labor-intensive and the costs are hard to trim. Many of the studies are labors of love. Many findings are esoteric. The worlds of commerce and social service benefit all too little from investments in these formal studies. The return may be greater for those who study their own shops and systems by these methods, but self-study so seldom brings the disciplined interpretations of the specialist into play.

Many qualitative studies are personalistic studies. The cares of observed human beings insinuate into issues of the present research. Privacy is always at risk. Entrapment is regularly on the horizon, as the researcher, although a dedicated noninterventionist, raises questions and options not previously considered by the respondent. A tolerable frailty of conduct nearby becomes a questionable ethic in distant narrative. Some of us "go native," accommodating to the viewpoint and valuation of the people at the site, then reacting less in their favor when back again with academic colleagues (Stake, 1986).

"It is not that we need more than a single indicator; it is the idea of indicator that is insufficient."

It is not simply a matter of deciding whether the gains in perspective are worth these costs. The attraction of intensive and interpretive study are apparent, and were earlier when qualitative designs were considered unworthy of respect by many research agencies and faculties—as by some, they still are. Researchers inquire. They are controlled by the rules of funding and their disciplines, but those influence how they will report their use of qualitative methods. All researchers use them. There are times when each researcher is interpretive, holistic, naturalistic, and uninterested in cause. Then, by definition, she or he is a qualitative inquirer. Administrators, too, have these leanings and use these methods. The question here is how disciplined concentration of these methods might improve the evaluation of systemic effects.

A Qualitative Strategy

Human Surveillance of Policy- One implication of qualitative methodology is to raise a caution flag on the use of "indicator variables"; yes, on all formal representations of complex phenomena. More than an intensive search for the closest indicator of an expected effect, we need disciplined scrutiny of this particular notion of effect. Interested in the effects of a research program on public policy, we may seek already-existing traces and we may create new indicators of changes in policy, but we should also extensively and repeatedly examine our conceptions of the research program and the public policy. Experimentalists (Boring, 1950) used to call it, "the criterion problem," the suitability of the representation.

As we first identify a program and a criterion policy, almost immediately we have expanding conceptualizations of the problem, the remedy, the effects. We have no single construct to represent, no true substance to indicate. It is not that we need more than a single indicator; it is the idea of indicator that is insufficient. We evaluators need to realize that we are asked for, and we ourselves yearn for, artifice, the hypothetical, the illusory. We propose indicators of things that do not exist other than in our imaginations. Many of the things we would indicate—the well-being of a child, the coherence of a curriculum, the fiscal integrity of a school district, the merit of a research policy—do not exist other than as mental contrivances. They are not things we can approximate. There is no way that we can test the validity of such "representations."

That does not mean we should purge our thoughts of indicators. We have no choice. Words are indicators, photographs are indicators, memories are indicators. We cannot communicate without representations of both the tangible and the intangible. Of course we will have indicators, not only in common discourse, but in all means of technical representation. The big question is how we will treat our indicators. Particularly, will we set them up as approximates to imagined truths, as substitutes for human sensitivity, for decision making? Will we use them to regulate our affairs?

Sometimes we will. We use various servomechanical systems: thermostats, cost-of-living increases, sliding scale cutting scores for admission. All, we hope, are subject to petition and override, but they are a part of our human systems. Some serve us well. Sometimes we wonder if they serve us well enough. The more the decisions impact indirectly on personal well-being, on differences in privilege, on the common good, the more we should worry that the indicators may be unwell and the more we should insist upon calibration in the form of close human surveillance.

It sometimes is supposed that a qualitative approach is fundamentally an aggregation and quantified analysis of data gathered in an qualitatively interpretive fashion (Miles and Huberman, 1984; Yin, 1989). While that may be useful, an essentially qualitative strategy for monitoring the effects of research is typified not by the establishment of quantitative indicators of qualitative phenomena, but by the establishment of disciplined and reflective human surveillance over all indicators, qualitative and quantitative.

These humans, these discerning humans, will use existing indicators and construct new ones. They will use multiple indicators to reflect the complexity of the phenomena and different perspectives found among people affected. They will couch their thinking and presenting of indicators in the language of experience, frames of time, place, and personality. If they do their work well, they will be a deterrent to overinterpretation of indicators, to the oversimplification of problems and solutions. They will demystify.

But they also will mystify. They will try to convey the best of insights of those who have most closely studied the matter. They will introduce new constructs, new models, and new scales. If they do their work well, they will not make it easier to command understanding, nor to make decisions. What they will offer is not indicators but virtuals, representations not of something real but essences of things understood. They will continue to remind us of the construction of our knowledge.

Interpretation Roles- Of the three pervasive characteristics of qualitative research I elaborated earlier, the most promising for extending NSF program evaluation is, I believe, interpretation. Interpretation is not a stranger at NSF, but more comprehensive and protected roles can be imagined. To come to understand the effects of major NSF programs, the qualitative strategy I propose is simple: an invigoration of interpretive responsibility, a mobilization of interpretive assets, an elevation of interpretive capability. I am echoing the plea of Cronbach and associates who called for much more vigorous collegial review of evaluation research (1980). The National Science Foundation needs comprehensive interpretation of what its science education programs are accomplishing (Katzenmeyer, 1993). The best contribution of qualitative methodology to such evaluation would be, I think, to enhance the role of systemic interpretation.

Individual evaluation studies aggregate poorly (Cronbach, et al. 1980), in

NSF as elsewhere. Policy makers do not get the support they need. Program officers and individual evaluation contractors provide too little in the way of historical perspective and independence. To get independent views of quality, evaluators are sought who have little to gain or lose by the conclusions they draw. These people usually have but cursory knowledge of present and past operations. To enrich formal evaluation with existing knowledge of present and past operations, an evaluation assignment often goes to prior funded parties (and potentially future award winners) or their associates, but these people are pressured by personal and institutional relationships to constrain their inquiries. There are natural constituencies of researchers for curricular issues, technical advances, teacher training, and special pedagogies, each capable of providing traditional reviews of research, development, and evaluation studies, but more narrowly defined than the panoramic responsibility for science and mathematics education. Most advisory panel members lack the purview, independence, and time to provide historical perspective.

"My suggestion here is ... for one group, an institutional council, to review science education performances of importance to NSF, including the systemic effects of its programs."

An Interpretation Council- One possible move would be to create within each NSF program or in the agency as a whole, an Interpretation Council, a small, full-time, internal but independent, evaluation-oriented policy-analysis team. Among the members should be persons well experienced in program evaluation, research integration (Cook, et al., 1992) and qualitative field study (Strauss and Corbin, 1990). Maintaining knowledgeable but dispassionate status would not be easy. Interpretation roles and council status would take time to develop. Although the appointments might be as difficult as those to the Supreme Court, the needed talents already exist among those who staff the Education Directorate. Members should be committed to gaining a thorough and enduring acquaintance with key issues, major projects, and related programs, yet having little vested interest in particular ones. This council should not replace the External Expert Panel, a more removed group needed for their interpretations (Katzenmeyer, 1993).

On the organization chart, the council perhaps should be a permanent free-standing affiliate, possibly attached to the Inspector General's office. Although much smaller, in some ways it would mimic the Government Accounting Office. GAO serves the Congress; the Council would serve an NSF program—but to provide interpretation and review rather than to complete studies. Like GAO, the Council should be obligated to stay relevant to the sweep of institutional responsibility, subject to multiyear mission renewal, and free to design and conduct individual program reviews. Even though dedicated to its sponsor, the Congress, GAO appears to me to have sufficient independence for designing studies, for occasional unwelcome findings, and for initiating some inquiries unrequested (Chelimsky, 1987). With strong management and a capable staff, I would say that presently GAO is the outstanding program evaluation shop in the world today. GAO is not an ideal model, however, because it does not maintain a sufficiently enduring relationship with particular programs. The purpose of that agency is not long-term administrative reflection and continuing program evaluation.

Thomas Cook (1978) and Lee Cronbach and associates (1980) pointed to the desirability of "social problem study groups." My suggestion here is similar but different. It is for one group, an institutional council, to review science education performances of importance to NSF, including the systemic effects of its programs. One organizational model to examine would be the fiscal audits provided by such corporations as Booz, Allen, and Hamilton. The audits are expected by both parties to resume annually until either party is no longer satisfied with the arrangement. Many of these auditing agencies have increased their staffing to offer program evaluation services. But here, too, there is little expectation that the persons working on the evaluation in a given year will have done so in the past and will build upon historical perspective. The format of the review is standardized to lessen the need for situational study. An interpretive council drawing from qualitative research methods would give particular attention to evolving situations.

The question may not be so much a matter of longevity of acquaintance as its intensity. Various corporations employ organizational and fiscal specialists to reside within the headquarters or plants for extended periods of time with a rather broad responsibility for discerning what is happening. When General Electric acquired the National Broadcasting Company in 1986, viewers were switching from the networks in great numbers to watch cable channels. Concerned about keeping the network profitable (Auletta, 1992) new Chief Operating Officer Robert Wright brought in a consulting team of four accountants to find ways of reorienting NBC away from revenue enhancement toward cost containment. GE officials wanted them to study "the culture" of the organization, which, through lengthy interviews, observations, as well as document review, they did. What the team provided were not indicators but hugely subjective estimates of what might be saved. They described the contributions of long-time NBC officials, especially those more bent upon providing public service than maximizing shareholder profit. The advice of the consultants was appreciated by corporate managers and disparaged by program staffs—but their interpretations were considered typical of what disciplined, intelligent observers will ascertain when they have sufficient opportunity to study a massively complex situation—not necessarily right but better than what was known before.

A long-staying internal but independent Council could be just as irrelevant as brief visitors and just as constrained as an internal team, but steps could be taken to increase relevance and minimize constraint. The Council could be guaranteed access, obligated by contract to raise critical questions, and insulated in various ways from intimidation. Such functions might be refined by the study of biographies of unique advisors such as Averill Harriman, Oscar Davis of the former U. S. Court of Claims, and Sam Messick of the Educational Testing Service. The Council could use its own internal workings to challenge observations and interpretations. In touch with principal investigators and evaluators, it could try out draft language and preliminary findings on program officers and other administrators. But mainly, it would serve as critical memory in the service of, but not dependent on, the science education program managers of NSF.

Drawing on the Qualitative Tradition- Whether or not an Interpretive Council is a good idea, the strategy of increasing the interpretive resources of the National Science Foundation should be considered. The present NSF investment in design of evaluation studies far outweighs its investment in interpretation. I have offered caveats here to recognize the shortfall in efforts to build a rational evaluation enterprise. I have presented my argument here in terms of the epistemological flaws in evaluation data and indicators that might be used to define the effects of Foundation programming, claiming that the usual indicators of need, productivity, or systemic effect are largely hypothetical, created more from social theory and political discourse than from empirical science. These indicators belong to a largely fictitious world referred to here as virtual reality.

It is within the capability of the educational research community to improve the present battery of indicators, from the Wallchart on up, but the utility of indicators appears to be to enhance or justify decisions already made on political grounds (Lindblom and Cohen, 1979). Rather than develop and validate better indicators, as many qualitative and quantitative researchers would urge, my recommendation has been to increase the quality of interpretation available to program officers, central administration, advisory panels, and oversight committees. Much depends on peer review, and peer interpretation, not just those peers on a special council, but all Directorate members. According to Michael Scriven (1992):

> Like democracy, peer review may be a flawed system but, if given its best possible implementation, it's the best in sight and something like it will always be a key element in proposal and program evaluation.

The emphasis in this paper has been not on review of projects or proposals but on review of program performance. Such interpretive evaluation could be accomplished in various ways (with the 1978 advice of Cronbach and associates still highly pertinent) but probably not with major reliance on external contracting and rotatory personnel. Institutional restructuring is needed—bringing greater disciplined interpretation inside. That needed interpretation, comprehensive yet program-specific, would require enduring study under security enjoyed by judges and scientists. I think the most important contribution the qualitative paradigm can make to the evaluation of systemic effects is to raise the emphasis on disciplined interpretation.

Informal evaluation of systemic effects of NSF programs already takes place; more formal evaluation is said to be needed. These programs are part of a political process and their evaluation is part of that political process. Efforts to shelter the evaluation from political pressure are needed: they cannot expect to be entirely successful. The qualitative strategy of increasing personal interpretation responsibility in a formal evaluation effort requires long-term agreements and protection to those who will bring bad news. A pressure-free environment is unrealistic, and explanations by interpreters are another form of virtual reality. But validation, experiential as well as technological, can engage the merely virtual in improving understandings of program merit and worth.

References

Aaron, H.J. 1978. *Politics and the professors: The Great Society in perspective.* Washington, DC: Brookings Institution.

Auletta, K. 1992. *Three blind mice.* New York: Vintage Press.

Ball, D.L. 1992. *Implementing the NCTM standards: Hopes and hurdles.* Paper prepared for the National Center for Research on Teacher Learning, Michigan State University.

Barnouw, E. 1970. *The golden image.* Oxford, England: Oxford University Press.

Boring, E.G. 1950. *History of experimental psychology.* New York: Appleton-Century-Crofts.

Chelimsky, E. 1987. The politics of program evaluation. *Society 25,* no. 1 (November/December).

Chelimsky, E. 1991. The politics of dissemination on the Hill: What works and what doesn't. Paper presented at the Conference on Effective Dissemination of Clinical and Health Information, 22 September, at the University of Arizona.

Cook, T.D. 1978. Speaking for the data. *APA Monitor* 9, (3).

Cook, T.D., Cooper, H., Cordray, D.S., Hartman, H., Hedges, L.V., Light, R.J.; Louis, T.A.; and Mosteller, F. 1992. *Meta-analysis for explanation: A casebook.* Russell Sage Foundation.

Cronbach, L.J., et al. 1980. *Toward reform of program evaluation.* San Francisco: Jossey-Bass.

Denzin, N.K. 1989. *The research act.* 3rd edition. Englewood Cliffs, NJ: Prentice-Hall.

Denzin, N.K., and Lincoln, Y. 1994. *Handbook of qualitative research.* Newbury Park, CA: Sage.

Eisner, E. 1979. *The educational imagination: On the design and evaluation of school programs.* New York: Macmillan.

Erickson, F. 1986. Qualitative methods in research on teaching. In *Handbook of Research on Teaching,* ed. Merlin C.Wittrock. New York: Macmillan.

Geertz, C. 1973. Thick description: Toward an interpretive theory of culture. In *The interpretation of cultures,* ed. Clifford Geertz. New York: Basic Books.

Guba, E., and Lincoln, Y. 1981. *Effective evaluation.* San Francisco: Jossey-Bass.

Guba, E., and Lincoln, Y. 1982. Epistemological and methodological bases of naturalistic inquiry. *Educational Communications and Technology Journal* (Winter): 232-252.

Guiton, G., and Burstein, L. 1993. Indicators of curriculum and instruction. Paper presented at the AERA annual meeting, Atlanta.

Katzenmeyer, C. 1993. Addressing program evaluation in federal mathematics, science, engineering and technology education programs. Unpublished paper. National Science Foundation: Author.

Krueger, M. 1983. *Artificial reality.* New York: Addison-Wesley.

Lindblom, C.E., and Cohen, D.K. 1979. *Usable knowledge.* New York: Basic Books.

Miles, M.B., and Huberman, M.A. 1984. *Qualitative data analysis.* Newbury Park, CA: Sage.

National Council of Teachers of Mathematics. 1989. *Curriculum and evaluation standards for school mathematics.* Reston, VA: NCTM

Polanyi, M. 1969. *Knowing and being: Essays by Michael Polanyi.* Chicago: University of Chicago Press.

Porter, A.C. 1993. School delivery standards. *Educational Researcher* 22, 5 (June-July): 24-30.

Psotka, J. 1993. An exploration of virtual reality. Paper presented at the AERA annual meeting, Atlanta.

Schwandt, T. 1994. Constructivist, interpretivist persuasions for human inquiry. In *Handbook of qualitative research,* eds. Norman K. Denzin and Yvonna S. Lincoln, Newbury Park, CA: Sage.

Scriven, M. 1967. The methodology of evaluation. In *Perspectives of Curriculum Evaluation,* edited by Robert E. Stake. AERA Monograph Series on Curriculum Evaluation, no. 1. Chicago: Rand McNally.

Scriven, M. 1978. Evaluating educational programs: The best models and their relation to testing. Paper presented at the Second National Conference on Testing, CTB/McGraw Hill, September, 21-22. San Francisco.

Scriven, M. 1991. *Evaluation thesaurus, 4th ed.* Newbury Park, CA: Sage.

Scriven, Michael. 1993. Hard-won lessons in program evaluation. *New directions for program evaluation,* 55. Summer. San Francisco: Jossey Bass.

Shavelson, R.J., McDonnell, L.M., Oakes, J., and Carey, N. 1987. *Indictor systems for monitoring mathematics and science education.* Santa Monica, CA: Rand Corporation.

Spiro, R. J., Vispoel, W.P., Schmitz J.G.; Samarapungavan, A., and Boerger, A. E. 1987. Knowledge acquisition for application: Cognitive flexibility and transfer in complex content domains. In *Executive control processes,* ed. B. C. Britton, Hillside, NJ: Erlbaum. 177-99.

Stake, R.E., ed. 1975. *Evaluating the arts in education: A responsive approach.* Columbus, OH: Charles Merrill.

Stake, R.E. 1986. *Quieting Reform.* Urbana, IL: University of Illinois Press.

Stake, R.E. 1994. Case studies. In *Handbook of Qualitative Research.* eds. Norman K. Denzin and Yvonna S. Lincoln. Newbury Park, CA: Sage.

Stake, R.E., and Trumbull, D. 1982. Naturalistic generalizations. *Review Journal of Philosophy and Social Science* 7 (1-2): 1-12.

Strauss, A., and Corbin, J. 1990. *Basics of qualitative research: Grounded theory procedures and techniques.* Newbury Park, CA: Sage.

van Maanan, J. 1988. *Tales of the field: On writing ethnography.* Chicago: University of Chicago Press.

von Wright, G.H. 1971. *Explanation and understanding.* Ithaca, NY: Cornell University Press.

Woolley, B. 1992. *Virtual worlds.* Cambridge, England: Blackwell.

Yin, R.K. 1989. *Case study research: Design and methods.* Newbury Park, CA: Sage.

We've had several models of discussions this morning, and I am going to introduce you to a third model. I am also going to talk about two of the papers.

The papers I have been asked to discuss today are very different, as you have just seen. In one, Bob Stake looks broadly at the field of evaluation, notes its gaps and its failures, its distorted emphases, and its unresolved tensions, and tries to build an evaluation mechanism for NSF that could perhaps remedy these problems. Specifically, the paper speaks to the promise of qualitative research, to the needs for experiential understanding rather than explanation, for interpretation rather than a search for cause and effect, for the distinction of system patterns of information over time, and for the conciliation of historical perspective with independence (I guess you'd say "semi-independence," Bob. I noticed that changed in the evaluation function.) The proposal is for an invigoration of interpretive responsibility to be incarnated by a group of "semi-independent" evaluation researchers within NSF. The group members would do some evaluations, advise on others, and generally lend their research expertise to the improvement of agency evaluation information over time.

The second paper describes a particular method—cluster evaluation—and proposes it as one likely to be useful to NSF in addressing two needs that its authors, Zoe Barley and Mark Jenness, judge important in the evaluation field today: the need to account for and conciliate the use of stakeholders, and the need to structure evaluations to serve the primary function of improving the program.

So, one paper focuses on a particular evaluation method, the other on a broad approach to assessment. One emphasizes knowledge, the other stresses the program and its services, but both papers deemphasize the importance of attribution of defined outcomes. I read both papers with great pleasure and think them worthy of NSF's careful attention and reflection.

Cluster evaluation seems to me to be a reasonable way of achieving buy-in and consensus across what are often warring groups. It's less clear to me how findings could be developed from the analysis —again Bob Stake's point about the need for validation—and whether so complex a process would be both feasible and productive.

Bob Stake's paper, which is a sort of luminous meditation on the problems and joys of producing something like real knowledge through evaluation, brings some critical insights to the assessment of teaching and learning. Reading his discussion of the distinctions between quantitative and qualitative representations of realities, I was reminded of the passage in Gabriel Garcia Marquez's *100 Years of Solitude*, in which the town of Macondo loses its memory and is forced to put up signs reminding citizens of the names of objects and how to perform functions like milking cows. By the way, the first object for which a sign is made is called a stake, spelled S-T-A-K-E, and of course another sign tells people exactly how to go about milking cows.

It's true that signs and other "virtual" quantitative abbreviations cannot represent everything, but sometimes it's the best we can hope for. My own bias in looking at an evaluation function—that is, how it should be organized and what methods are most valuable—would add some other components to those presented in these two papers. To me, the kinds of evaluations that need to be done will always depend heavily on three things: the kinds of policy questions or evaluation questions that will be asked about the program, the service, or the function; who will be asking these questions; and what evidence will be needed both to answer the questions and satisfy the political and institutional culture of the particular audience. The question, after all, is the critical trigger that determines what methods need to be used.

Someone asked the question earlier, Can we really separate evaluation from dissemination?

[redacted]

Again, that depends on the question. If we are looking at something that the Congress might ask us to do—for example, evaluate a study and tell us whether it's good—we would simply do an evaluation of it. We would critique it in one way or another, depending on what the study was, but there would be no need for dissemination other than simply passing it to the committee that wanted it. If we are talking about a program where the question is, Can we use intermediaries to disseminate knowledge to a given audience? then dissemination is part of the evaluation—it can't be separated. So it all depends upon the question that is asked.

I think we shouldn't forget that traditional quantitative and qualitative methods can answer a great many questions about the effectiveness of programs or functions and the quality of services (for example, questions about whether someone learned something or not, or whether program beneficiaries are pleased with or insulted by the services they receive). But ingenuity and creativity and innovation are needed to answer broader, complex, systemic questions.

To me this suggests four interdependent means of dealing with these broader issues. The first is an evaluation organization that starts with a profound understanding of which questions will most often emerge, and why, from the political environment within and surrounding an agency and its programs. The second is a panoply of traditional methods and the skills to apply them appropriately and to validate them. The third is the exploration of new methods as a response to questions that cannot be answered with old ones, and the fourth, an in-house organization that can demonstrate the feasibility and usefulness of doing both the old and the new. New methods cost a lot of time and money to specify, test, and apply, and they involve some risk to their users. In particular, the more political controversy there is about a topic, the greater the initial credibility risks of newly developed methods. Therefore, the evaluative requirement for them should be, I believe, abundantly clear and their use warranted by the need for answers to specific user questions.

[redacted]

David B. Rymph - ACTION

I want to begin by expressing my gratitude to Zoe Barley, Mark Jenness, and Robert Stake. I want to thank them for giving me the opportunity to read their papers and learn from them.

My thoughts are organized into four themes. First, ideas, solutions, and innovations have difficulty moving horizontally in hierarchical systems. Second, local-level project personnel in social programs can do program evaluation, if technical assistance is available. Third, qualitative analysis is central to the evaluation process. And fourth, NSF needs to study the problems of math and science education in a larger social context.

Promoting Horizontal Flow of Information

I have a lot of experience working in local-level programs, and I have learned that information usually flows vertically in any institutional system. Reports, plans, audits, monitoring results, evaluations—all of this stuff moves from program units through management to policy people. Few resources are given to moving information between program units. Consequently, the people who are responsible for delivering services in a program rarely have means or opportunity to communicate with each other.

Cluster evaluation, as described by Zoe Barley and Mark Jenness, does much to overcome the horizontal flow problem. In the cluster approach, regular networking conferences for the projects are a central feature. Staff from different projects participate in negotiating agreed-upon common outcomes and then collaborate in data collection. Finally, "dissemination of findings and sharing of lessons learned occurs between individual projects in the cluster..."

Local-Level Evaluation

In my current job as Director of Program Evaluation for ACTION, the Federal domestic volunteer agency, I been actively engaged with the problem of how to get project staff involved in evaluation. My agency gives grants to community-based organizations. Many of those grants carry a congressionally mandated requirement that they conduct an annual evaluation of their programs. For small grants, say under $100,000, this may appear to be an absurd requirement. The resources needed to meet the evaluation standards of the grant guidelines are seen by project personnel as detracting from their basic mission, which is not research. In small programs, often the evaluation tail is wagging the service delivery dog.

Through ACTION training conferences for grantees, I have made some efforts to overcome this problem. I try to give project personnel some skills in what I call local-level, low-tech, low-cost evaluation techniques. For example, I ask participants (and sometimes I might have a few hundred in a room with me at one time), "How many of you know your annual budgets?" Everybody raises a hand. Next, I ask, "How many of you know how many hours of volunteer service your project produces each year?" Almost everyone raises a hand. Finally, "How many of you calculate the cost per volunteer hour of service?" Rarely have more than 3 or 4 persons in 100 responded affirmatively.

Again, cluster evaluation proponents recognize this problem and opportunity. The cluster evaluation approach emphasizes the central involvement of evaluation in program management and improvement and stresses the importance of direct stakeholder involvement in that evaluation. The processes of cluster evaluation, as described by Barley and Jenness, go a long way toward empowering local-level project people with needed evaluation skills and other resources.

Qualitative Analysis

In reading Bob Stake's paper, I was reminded of a time years ago when I was doing extended field research in Johnson County, Kentucky, the birth place of Loretta Lynn. In some of the Pentecostal churches in that part of eastern Kentucky, there was the belief that a person possessed by the Devil could not say the word, "J-, J-, J-Jesus!" Well, Bob Stake apparently is possessed by some demon for he cannot say the word "A-, A-, A-, Anthropology!"

He refers to several concepts and methods that are the traditional domain of cultural anthropologists. These include ethnographic research, the emic/etic distinction, and holism. In one passage, he presents a fair representation of anthropology's central concept, culture.

We have common knowledge not because there are pre-existing facts—truths—for us to discover, but because learning, just like dressing and driving, is a social process. We have strong tendencies to conform. We modify our actions to fit the actions of those we respect. And we create knowledge very similar to that of the people around us.

Stake mentions several of the social sciences, but nary a mention of the father and mother of qualitative research, anthropology.

I recommend to this audience the extensive research literature in applied anthropology. In this subdiscipline of anthropology, the concepts and methods that Bob Stake discusses are not nontraditional, rather they are very central to our tradition.

One caution: qualitative research is not easy. Bob Stake is absolutely right in characterizing it as costly, time consuming, and subjective. My experience with contractors conducting research for my agency may be typical. Our research designs often call for site visits, case studies, and other types of participant observation. I have yet to see the wealth of information gained in these qualitative approaches used in any way other than as anecdotes to fill out quantitative reports.

I would disagree, however, with his contention that the "results pay off little in the advancement of social practice." While a reply would need another paper, I must say that applied anthropology has made major contributions to improving social conditions, especially in the developing world. One example is the important role that anthropological (qualitative) research is playing in the development of techniques to disseminate health information on AIDS in Africa.

The Larger Social Content of Math and Science Education

As a final comment, I want to suggest to the National Science Foundation that it expand its research on the problems with math and science education in the United States. In addition to improvements that might be made to curricula, we need more understanding of the cultural settings for science education in our country.

While we are a nation that seems to revel in technological advances, we are also a nation beset with rampant superstition, ignorance, and even rejection of basic scientific processes, principles, and theories. Almost a majority of people in this country, if some recent polls are to be believed, accept the creationist view of our origins (the story in Genesis) and reject basic evolutionary theory. Millions profess to believe in astrology. The list of irrational belief systems that are being embraced by substantial numbers of Americans is quite lengthy.

The question for NSF is, How can we educate children in science when their parents show such disregard for its most basic principles?

Overview

Michael Scriven

I want to begin by saying how important I think meetings like this are, that is, meetings in which the existing paradigms of evaluation are seriously questioned by those who are not only involved in the game, but also those who are hiring these people and those who are being evaluated by these people. I think we should regard it as a kind of moral imperative for evaluation as a discipline that meetings like this happen.

The results of the major efforts that we have heard about today are impressive. One of the results is a series of suggestions on a very practical level, in particular, a list of 40 suggested questions that you might ask in doing an evaluation of programs like the examples from NSF. There is no substitute for the local experience that some of these people have as evaluators and as program participants. While their comments are aimed at NSF many of them will work equally well for another agency. Many are generic types of questions, though specific enough to be relevant to the ground level of evaluation. So, I think simply on that ground alone, we have something worthwhile here.

On the other hand, there was, I thought, a substantial lack of clarity about what was being done in the efforts discussed today. That doesn't mean that they're not useful. It's just that the interpretations given them were sometimes implausible.

The three things that were going on in these papers, apart from trying to improve evaluation, were:

- Trying to improve dissemination;
- Trying to improve explanation and understanding; and
- Trying to improve description of process—what happened? How did it come about?

These three things need to be distinguished, not sharply—that's not possible—but generally speaking, as carefully distinguished as possible. I think we are meant to be talking about evaluation. Let me put it another way. Dissemination is a specific process that's crucial in certain projects, but absolutely irrelevant in others (e.g., where you are trying to solve a theoretical problem, and the payoff is having solved it). The justification for the project is that it had a reasonable chance of solving the problem, not that it did solve it. Dissemination, as Eleanor Chelimsky put it, is going to come in if the task of the evaluation is to find out whether the results were disseminated successfully, and it's not going to come in if the task of the evaluation was to find out whether the problem had been solved, useful discoveries had been made, etc. I think this distinction is quite unclear.

"The results of the major efforts that we have heard about today are impressive."

One of the reasons for that lack of understanding leads to a constructive conclusion that we should take extremely seriously. We really are not treating dissemination as a research area, although it's very unfortunate that we are not. We're constantly reinventing wheels, or much worse, we're starting to realize that someone already did, but we don't know how. There are lots of tricks out there in "dissemination land," and even some experts in some parts of it, as you well know. But we're not treating it as a body of knowledge we must have to get many of our tasks completed.

Dissemination is, of course, a perfectly sensible part of applied social science. We just need to give it more attention and expect to get more from it. Then, we can pull that knowledge in without having to force people who want to help in changing the schools to be experts on dissemination, which many of them are not.

"I do not have the faintest understanding, nor does anybody else, of why aspirin works. But as an evaluator in the pharmacological field, it's not a big problem to prove that it does."

The explanation and understanding issue is a little trickier because there is a gray area. Bob Stake spent quite a bit of time talking about the importance of qualitative research as a way to obtain insight and understanding, perhaps on the way to explanations of certain kinds. Well, there is a part of evaluation where explanation and understanding is of the essence. It's what you might call "perspectival" evaluation, where what you are doing is trying to achieve a new perspective on the program — to see it in a different way. Wittgenstein spent years toward the end of his life working on the phenomenon of seeing one thing as another thing. That's a very important part of what the good evaluator, and particularly a good qualitative evaluator, can do. But it's only part of the job, and it's only part of the job in some kinds of evaluation tasks. So, we want to be careful about thinking that explanation and understanding is, in general, part of the evaluation job. It is not. I do not have the faintest understanding, nor does anybody else, of why aspirin works. But as an evaluator in the pharmacological field, it's not a big problem to prove that it does. I don't want to be fooling around too long with people who keep saying, if you can't understand how learning goes on, then you can't evaluate teaching. Of course I can evaluate teaching; I don't need to know anything about learning theory, I just need to be able to recognize effective teaching when it bites me.

So we don't want to get into this academic trip about the need for the theories in order to do good evaluation. On the contrary, in many cases, if you can't do good evaluation, you can't even develop the theories of good teaching. Evaluation is the groundwork without which you cannot validate the theory. You want to know what methods of teaching work better, so you need to have measures of learning, not the theories of learning, which you can use to find out which methods did work better. You must be able to evaluate the learning, assess the students' work in order to evaluate the theories.

Indeed, there was one paper which was almost entirely devoted to discussions of questions about how things happened (i.e., descriptive research on various processes in learning and teaching). That's important stuff, it is a part of the task of the RTL program, but it's not a part of the task of evaluating teaching.

There were thus four kinds of valuable payoffs from the papers and comments. First, we were presented with a wonderful array of suggestions for indicators and questions to be asked when evaluating important programs of this general type. Second, there were suggestions for needed research on evaluation. Third, quite a different matter, there were suggestions for research on how teaching and learning works. And some of the suggestions for research on evaluation were, in fact, suggestions for research on dissemination or research on explanation. We can shuffle those over to other groups where they are useful topics for research, but not of direct concern for use in evaluation here.

Fourth, there are the proposed "new models," and one aim of the conference was in terms of looking for new models or approaches. Here I think the arguments are less persuasive, and I find myself in the truly embarrassing position of defending the status quo, something which I've never done throughout my life. But there doesn't seem to be anybody else around to say, "Hey, that's a straw man, we do better than that today." So I'm going to argue in that direction for a while.

We need to distinguish first between the arguments that we do need a new approach, and specific arguments for the proposed new approaches. We have heard quite a few of both of these. The arguments for needing a new approach are, in my view, mostly aimed at what is really a straw man. Now, NSF has had a great deal of experience with the standard approach to evaluation because it sends out a lot of RFPs to get evaluations done and it sees what comes in. So I'm not going to second guess their view, that there's a body of bidders who trot out their favorite quantitative something or other model. Yes, things creak at the joints a bit in the process of development, but one doesn't really want to treat that as the state of the art. If we're going to start looking for new paradigms, then we need to see if the existing best practice is faulty. And the best practice isn't always what Brand X trots out with their number 16 proposal writer when you run an RFP up the flag pole. Best you can get from Eleanor Chelimsky; the best you can get from the best of the audit agencies; practice is the best you can get from the best of the OIGs; the best you can get from the best practitioners in the American Evaluation Association none of whom are bidding on these RFPs. We want to be careful that we don't rush to ditch current best practice on the grounds that current proposals are unsatisfactory for the sort of tasks that are involved in evaluating the types of programs exemplified, but not restricted, to the three big NSF programs that were mentioned frequently.

It seems to me, for example, that the best current practice is a kind of eclectic amalgam of qualitative and quantitative. It's certainly not just quantitative. And this is not only for the reasons Bob Stake gives that there is no such thing as pure quantitative, but also for the other reason that these days best practice will have explicit qualitative elements aimed at various areas such as those where you can't get a good quantitative grip and those where the interpretive process is absolutely fundamental. Numbers aren't going to do the interpretation for you. So, it's an eclectic mix of quantitative and qualitative, formative and summative, internal and external, worth and merit. That is, it involves looking at cost effectiveness and not just effectiveness.

"We need to distinguish first between the arguments that we do need a new approach, and specific arguments for the proposed new approaches."

In the *Call to Arms*, Joy listed the reasons that the Directorate had for suspecting that there might be a need for a new paradigm. She says that the traditional approaches are "not directly applicable to the many research-oriented, ground-breaking, inquiries" that NSF often supports. Well, of course, "ground-breaking" is an interesting phrase; it does suggest that you broke some ground. And, if you broke some ground, it does suggest that there ought to be some sort of a footprint in the sand. We should at least see some sort of a new path, some blockage that got broken through, some problem in conceptual understanding was solved. So I don't feel that we really should have to say, "Abandon hope all ye who enter the eclectic, contemporary model of evaluation," here's a case where you can't handle the challenge. Groundbreaking is easy; at any rate groundbreaking is a lot easier than, "Did it have an effect on the kids in the 12th grade in the United States?"

The research efforts in RTL, for example, are in an important sense, much easier to evaluate in themselves. However, the question of whether everybody has come to recognize the leading work in the field, whether the practitioners have all been affected by these efforts, is the dissemination question. It involves another step, and it's harder. The question of whether the problem is solved is not so hard. And so, I think it's a really serious reason for avoiding naive applications of a quantitative model, which you certainly run a risk of getting when you put out RFPs. But you should expect to write your RFPs to rule out the naive bidders, expect to be very tough about awarding contracts, and restrict awards to people who see through the simple-minded ways of handling the issue at hand.

Joy adds that the impacts are different between studies that are research oriented and those that are groundbreaking. For example, she says that the old style of ground-breaking evaluation "seeks to attribute the effect to a single source." Well, is that really true? They were interested in the question of whether somebody's project did it, if that is a single source, because that's what they were asked to find out. But, then you can hardly blame them because they looked at the question of whether a single project did it. I find myself wading through many pages of their variance analysis, which says, 'No, there isn't a single source that did it, but the single source contributed something to it, here's the figures to prove it.' That doesn't seem totally stupid to me; it seems to me that's a fairly sensible kind of approach. So I think we can handle the notion of more than one source, and even the quantitative fellows, bless them, actually do that quite a bit, and certainly the rest of us can do it too.

The second thing she says and, of course, Joy didn't invent all this out of whole cloth—she's picking up common comments—is that standard evaluations are almost entirely reliant on quantitative data. Well, that is a sign of weakness in the bidder, in the evaluator. Let's not make any mistakes about it, if they're almost entirely reliant on quantitative, then in very many cases that will be just a flaw in their capacity to solve the problem of getting a true measure of merit and worth. But that seems to me to be an example of bad use of a simple-minded paradigm, not an example of current best practice being unable to handle the problem.

Following up on this point, she says that quantitative won't do because a single successful project may justify the entire research investment. Indeed, but where do we have somebody saying the program was a failure because only one Einstein went through? Nobody says that; or if they do, then scrape them off the list for next time around.

We can cope with selecting portfolios of high-risk, high pay-off investments. At the first meeting of the Evaluation Network, 20 something years ago, I set that task as the task for the President's prize. Nick Smith won the prize for a study in which he showed how to handle portfolio assessment. It's discussed in some references as the apportionment problem. So, we want to be careful about hopping on a new bandwagon on that issue. I'm speaking reluctantly in favor of the existing best practices being better than you might think.

One of the things I do at the moment is handle all the external evaluations for a wealthy community foundation that funds absolutely everything you can think of—legal aid, work in San Quentin, housing for dispossessed mothers, help for the drug addicts, restructuring schools. Mention anything, we've got a program, probably six. Now, that's a very wide variety, but we don't find any need to shift paradigms among them. In fact, the value of somebody handling a wide variety of evaluations for the trustees of the foundation is that they can use a consistent model across the board. It gives them a degree of comparability which is useful. Perhaps we ought to think the same way about large agencies. We should be trying to use a standardized model—which doesn't mean a primarily quantitative model—across the board.

Then there was the question of the tendency to give priority to measures of student achievement. Well, is it an inadequate sole measure for some NSF programs? Certainly it is, and if you were to use that as the only measure, you would have to wait around 25 years to get some of the data, which wouldn't be much good.

So the real rival for the new style religion is the reformed orthodox church, not the church of the 1960's. Bearing that in mind, we now come to look at the proposed new models. These are not very much of a threat to the reformed orthodox model; they are much better seen as suggestions which should be used to forge refinements of the eclectic best practice model. I think that they can be very useful in that role. Cluster evaluation for example, seems to me an excellent device for improving evaluation, if we redefine it. Redefined, it looks something like this. The evaluation staff, on a group of related projects, regularly meet to discuss what they are doing and how things are going; and occasionally, but only occasionally, meet with the project directors in order to discuss how things are looking, but in limited terms, not full disclosure at all. In the way in which this was described to us here, it was really a replay of the original, transactional, North Dakota, East Anglia, model of collaborative, negotiated evaluation. Which, to put it bluntly, is a great way to cheat the consumer. Who's represented at the negotiations? It's an exact analogy to the way in which the union meets with school district management to thrash out the contract. Who's not there? There's nobody representing the kids, nobody representing the taxpayer. And you get just the same amount of credibility with the results. So, in this case, getting the project people in bed with the evaluators is exactly what you do not want to do if you want a credible and serious evaluation. Now, that approach is very popular these days; the President of the AEA calls it "empowerment evaluation." But it's simply a way to guarantee the loss of what objectivity is possible in those ongoing, formative evaluations, and that's a terrible loss. Why do you read Consumer Reports? Why don't you just read the handouts from General Motors? Well, suppose we insisted that the Consumer Reports auto evaluators spend the year with GM engineers. Will that improve the objectivity? No, it will corrupt it. We knew that from day one. So, I don't feel happy about that example.

"... getting the project people in bed with the evaluators is exactly what you do not want to do if you want a credible and serious evaluation."

It seems a bit mean to have picked on the cluster evaluation protagonists and then not to go pick on everybody else, which I could easily do. But instead, I'm just going to do two remaining things. First, I'm going to put forward what Bob Stake will regard as a truly straightforward demonstration of my simplemindedness, by defending the silver bullet approach. Then I'm going to talk about Bob Stake's paper.

Now, I'm going to ask you in thinking about this intervening discussion where I want to convey to you, what I believe we ought to be doing, to think of three people. The first is Mosteller, whose name was mentioned earlier. Fred Mosteller at Harvard is generally thought to be one of the two or three best applied statisticians in the world. He's the author of *Understanding Robust and Exploratory Data* which was a reality-oriented push in statistics. He is also the author of another notion which I want to commend to you today because I intend to use it as a paradigm. After years of editing a journal and receiving countless submissions in which something or other turned out to be statistically significant at the .05 or the .01 level, he coined the term, "interocular differences" to contrast with "statistically significant differences." His line about them is very simple. Go ahead and play around with the statistically significant differences while you are doing research because it may help you find something interesting. But don't come to me until you've found some interocular differences. In other words, if the difference doesn't hit me between the eyes, I don't want to hear about statistical significance. Now that's the voice of a good statistician and it's a very sensible appropriate voice when you look at what happens to the 95 percent of published research that was statistically significant. It doesn't replicate the second time around, it turns out to be trivial in the light of various conditional requirements on it, and so on, and so on, and so on. So the first point is, we ought to be looking for interocular differences in evaluation and we ought to be sending the statistically significant stuff back to the drawing board.

Now, the second person I'd like you to keep in mind, though you haven't ever heard of him, is John Hattie. You'll hear a lot about John Hattie. He's a brilliant eclectic educational researcher, my fellow professor at the University of Western Australia for several years. He's done an analysis of the kind that will make Bob Stake want to bring his lunch back, the kind of study which Congress just loves to get. It's this. He's looked at every educational intervention that can be given a generic description, such as should we add paraprofessionals; should we put computers in the classroom, in what ratio; should we reduce class size; should we increase inservice education; should we mainstream; should we ability group. He simply lists them, and does a meta-analysis, or finds another meta-analysis that has already been done on each of them. He finds the effect size and lines it up, and he says, if you've got X bucks you can possibly spend in a school district, here's the shopping list in order, this is what you'll get for each buck.

You'll remember that Hank Levin has done a very nice study of that kind, aimed particularly at whether you should computerize or not but covering other things. Hattie has a generalized version of that. Of course, this will not be a perfect guide, but as Bob Stake says, we have to move from initially misleading indicators to better indicators. Now that's the kind of result that Congress is always pounding us for and that academics sneer at, but I think quite wrongly. In this connection, one should remember the story of the Office of Inspector General. There was one Inspector General 15 years ago, and there are 26 today. Why? Because the academics would never get the evaluation reports in until long after the people who needed them had left. An Inspector General finally said, I think it can be done in 3 months for $100,000, and so let's see. And, so now we have a whole bunch of people doing those evaluations. Have the academics ever done an evaluation study to show that these are such trashy results that they have lead to millions upon millions of wasted money? No, they have not. Now that either shows that they don't want to find out, or that the results aren't at least obviously disastrous. So, I think exactly the same thing applies here: meta-analyses should guide policy. We want to be very careful to try to speak the language of common sense on these things.

I'll bring that down to cases. In the Advanced Technology program there is a great deal going on, but in 25 years of serious work in the Ed Tech area, I have found the same problem to be endemic that I see in the material here, briefly described though it is. You might sum it up by saying that they'll never look at the top competition. If you're looking for magic bullets in the Ed Tech arena, you won't find them by test firing against bows and arrows. Magic bullets have got to be the ones that beat the best of the other bullets; it's not interesting that they can beat bows and arrows. And we're finding a lot of material here whose only claim to fame is that it can beat a bow and arrow.

Specifically, there's very little in Ed Tech that can beat a programmed text, but we never run things in Ed Tech against programmed text. We run them against the status quo, non-Ed Tech approach, or against very primitive Ed Tech approaches. That's not serious evaluation. Programmed texts have now gone: "everybody knows" that they don't work. But there were many out there that could beat anything. They could beat intensive tutoring, they could beat the best teacher there was, they could beat what existed then in the way of computer-assisted material. And, so we just walk past that; we averaged it out. Who cares about the average? The question is, what was the state of the art? Certainly programmed texts were more expensive than standard texts, but a lot less expensive than most Ed Tech. So, one of the problems that we've got, is that the group of Ed Tech folk, are, to put it bluntly, massively biased in judging proposals. What is the effect on them of using the toughest possible standard, competing against the best alternative there is? It is that very few of them will ever be funded. They know that very well, so that you must understand that a lot of what I have to say consists in saying, don't do collegial review, don't talk peer review, if by that you intend to mean people from the same in-group, because they are massively biased.

"There was one Inspector General 15 years ago, and there are 26 today. Why?"

Now, with respect to Bob Stake's final suggestion about a panel, I'll suggest how one might expand that notion, so that you would, in fact, get quite a good degree of independence. When you do a secondary school accreditation, it's always a bad deal because when the team of 40 arrive at the high school, it's got one person on it in Driver's Ed, and one person in Accounting, and one person in whatever, and after the Driver Ed person goes to look at Driver Ed and has tea with his friends he saw last week at the All-State Conference in Driver Ed, he then comes back saying, "Gee, this school is strong in Driver Ed." What's that worth? Nothing. If you'd sent the accountant to look at Driver Ed and the

Driver Ed guy to look at Accounting, we might have learned something. Better, send both to both. We should use that model for panel construction—the mix of local and outsider expertise.

So, remember Mosteller, remember Hank Levin on the employment futures that high tech delivers and on the relative payoff of various ways you can spend money on student outcomes. Remember John Hattie doing that more generally, and me talking about the programmed texts as the main competitor with CAI, e.g., with enormously expensive PLATO installations. I did the largest evaluation of a PLATO installation that's been done so far, so I speak with some interest in that area.

The bottom line of that sort of study, from Mosteller through Hattie, is the sort of thing that Congress rightly wants to see. Academic condescension says, 'No, that's a naive assumption about how easily you can produce indicators for these things.' I think not. I think the fact is, that we ought to revitalize the entire effort so that the task is this: using the Ed Tech area as an example we'll give you a little money for a pilot; then if you show signs that you can beat a programmed text, we'll re-fund you for a limited period of time. If we want magic bullets, we have to set the shooting competition up with the proper rules; beat the best, or go back to the drawing board.

"If we want magic bullets, we have to set the shooting competition up with the proper rules; beat the best, or go back to the drawing board."

Footprints: A Search For New Strategies For Evaluating EHR Programs

Laure Sharp and Joy Frechtling
Westat

Prologue

This paper presents our interpretation of what was said at the "Footprints" conference and written in the "Footprints" papers. It is not an attempt to summarize all suggestions or to comprehensively discuss the pros and cons of each author's proffered strategies. Rather, we have attempted to extract the points that we see as especially relevant to the Division of Research, Evaluation, and Dissemination (RED) and to offer our suggestions for how RED can build on what was learned from the "Footprints" task to shape its future evaluation agenda.

Introduction

In 1994 and 1995, several programs funded by NSF's Directorate for Education and Human Resources (EHR) are scheduled to undergo third-party evaluations. Planning these evaluations will be a complex task, given the heterogeneous nature of the programs and the projects that they support. As a first step in the planning process, the National Science Foundation asked Westat to commission a series of papers from experts in diverse fields of evaluation to help develop a framework for examining these programs. The eight commissioned papers and the comments of seven discussants are presented in this volume. In this final paper, we have sought to highlight and discuss those topics and ideas that emerged from the conference and seemed most germane to EHR's planning needs. This selective review was guided by what we believe are EHR's concerns and especially those of RED in undertaking program evaluation in the near future. Many more valuable ideas and comments can be found in the papers and discussions, and they deserve close review by NSF staff and others interested in innovative evaluation practices.

The need for a New Evaluation Approach

New techniques were sought because the RED staff felt that traditional educational evaluation methodologies would not be appropriate to assess what many EHR programs had accomplished.

Traditional evaluations of educational programs have been developed primarily to assess the results of new or improved service delivery models. For example, Chapter 1 and Headstart typify the service delivery model and provide the template against which most large scale federal evaluations have been constructed. In such evaluations, typical questions include the following:

- Do students benefit from the introduction of new services or technological innovations, such as the use of computers?
- Do students' attitudes, interests or test scores change?
- Do teachers adopt new instructional methods after attending science workshops?
- Do these new methods result in improved student performance?

The service delivery model may be appropriate for some EHR-funded projects. However, it is ill-suited to many others, and with a few possible exceptions, it is inappropriate for the evaluation of programs. The mismatch stems from a number of sources, including the organization and makeup of the EHR programs, the goals the programs are intended to meet, and the very nature of the funding mechanism that predominates.

Each of these is considered further below.

Program Structure

Traditional evaluations have been developed to assess the impact of programs supporting projects that are fairly homogeneous in nature. They have common components and may even be built along a "planned variations" model. EHR programs, including Research on Teaching and Learning (RTL), Applications of Advanced Technologies (AAT), Studies and Indicators, in contrast, support a wide variety of projects that are highly diverse and vary in size and duration. Some are part of a stream of research, reflecting decisions made over multiple funding cycles. Some reflect the results of cross-program collaboration. Others are one-time efforts or exploratory projects.

While some of these projects can be evaluated using a service delivery model, for many others the model is unsuitable or, at best, incomplete. For one thing, it cannot be applied to projects that can be categorized as basic, theory-driven research (as contrasted with those categorized as applied, problem-based research). It is also inapplicable to descriptive studies and those that are funded by the Studies program to generate new international statistics on student achievement in mathematics and science (SIMS and TIMSS).

Even where the model may be applicable to individual projects, it is rarely appropriate for the evaluation of a program as a whole. That is, in many cases, it may be neither possible nor conceptually correct to aggregate individual project evaluations for the purpose of evaluating the program as a whole, if only because a comprehensive program evaluation must answer questions that go beyond assessing the outcome of individual projects. For example, to evaluate the RTL program, policymakers and other stakeholders may want to know if the funded projects addressed the most important research questions or had an impact on classroom practices in school systems other than those in the project sites. Aggregating the evaluations of individual projects does not provide answers to these more global questions. Furthermore, some programs - of which AAT is the prime example - may choose a "high risk - high gain" investment strategy, anticipating that only a few projects will lead to scientific breakthroughs. In this case, an evaluation based on aggregation of project outcomes would be especially inappropriate.

Program Goals

A second obstacle to using the traditional, service delivery model for many EHR programs is their broad-based and highly ambitious goal structure. Traditional evaluations have frequently been motivated by, and structured to address, specific legislative mandates. Rightly or wrongly evaluators have relied primarily on narrow goal specification and looked for indicators that can document goal attainment over a period of a year or two or even five.

The EHR programs on which we are focusing lack specific, tangible goals that are to be met within a given time period. While the ultimate objectives of NSF's programs in education and human services are clear, they are also very ambitious and very broad. The programs serve to promote more participation and better learning outcomes in mathematics and science among students at all educational levels and/or more recruitment into scientific careers especially for underrepresented populations. It is very difficult to assess progress toward these goals in the short time span under which program evaluations must typically operate. Further, given the magnitude of the implied task of changing major components of the educational system, holding the relatively modest NSF programs accountable for their attainment is unrealistic.

The Funding Mechanism

Perhaps the greatest obstacle to the use of traditional evaluation strategies for NSF programs stems from a third cause — the funding mechanism. Educational programs and projects for which traditional evaluations have been carried out were usually funded through contracts or grants that prescribed performance requirements, benchmarks, and outcome criteria. In the great majority of cases, EHR programs are based on the academic grant model, where grants are awarded to field-initiated projects selected through peer review. In this process the emphasis is on quality of performance and the qualifications of the principal investigator. Awards based on the academic model encourage experimentation with innovative ideas and processes; the grantor will, therefore, accept a high risk of failure as part of the research design. The process is tolerant of considerable deviation from proposed activities in the detailed execution of the project, at the discretion of the principal investigator, and gives investigators considerable leeway in their choice of procedures; adherence to specific performance criteria is seldom required. This grant model is in line with NSF's basic funding mechanism and philosophy for the support of research in the physical sciences.

As a rule, institutions using the grant mechanism to fund projects do not carry out systematic program evaluations. Rather, grant programs sponsored by government agencies and private foundations have relied for evaluation on judgmental approaches through expert panels, review committees, and similar mechanisms. Education programs are also being reviewed in this manner, but the mandated periodic third-party evaluations call for more systematic approaches.

Thus, RED must develop a strategy for the systematic evaluation of EHR programs whose goals and funding mechanism often preclude the use of methodologies traditionally used in the evaluation of education programs.

The Guiding Concept Proposed by NSF: Footprints

Understanding the difficulty posed by the need to evaluate many of EHR's programs, NSF staff sought new ways of examining program accomplishments. The "Footprints" model was chosen because it seemed to offer a new way of thinking about results and because it seemed flexible enough to apply to the evaluation of the very diverse programs funded in EHR.

"Footprints" were defined as evidence that the program had left a mark on the field of mathematics and science education and had contributed to new knowledge or new practices. Specifically, this metaphor suggests that the program evaluation should seek to ascertain whether a program has contributed substantially to the state of knowledge in mathematics and science education (the "research base"), and has left its own "footprints in the sand" (evidence that both researchers and practitioners have been exposed to this knowledge and/or have been influenced by it). A footprint implies that a mark has been left, but it is not explicit with regard to how and when the mark actually got there. This metaphor has the advantage of not being overly specific as NSF's Susan Gross said in her introductory comments, "Footprints come in all sizes and shapes," thus avoiding a priori restrictions on potential outcome indicators. RED staff initially identified four general areas where footprints might be found:

- Effects on the profession (the supply and characteristics of researchers, topics presented at conferences, and in journal articles);
- Effects on other research;
- Effects on practice (teacher training, curricula, and implementation of sound pedagogy); and
- Effects on funding agendas of other institutions.

Such footprints might begin to answer the broader questions which NSF itself, as well as oversight agencies within the Federal Government and congressional bodies, ask about these programs:

- What has been their impact on the thinking and practices of educators and administrators in local school systems?
- Are these programs likely to contribute to the achievement of national goals such as higher participation by women and minorities in mathematics and science education?
- Is there any evidence that they have improved the quality of instruction in science and mathematics at various levels of the educational system? Have the programs affected the thinking and actions of educational policymakers, of researchers, and of those who fund research at the national, state or local levels?

Ideas and Suggestions from the Conference Papers

As might have been expected, given the diversity in their backgrounds, work settings, and disciplinary orientations, each paper author and discussant came with his or her own experiences, approach, and ideas. While some presenters dealt extensively with the "Footprints" theme, others addressed the issue of nontraditional analytic techniques or, more broadly, the topic of nontraditional approaches to educational evaluation. As Joy Frechtling pointed out in her introduction to the conference, while none of the papers went so far as to propose a specific evaluation design for one or more EHR programs, they provide valuable directions and inputs. Many of these can provide useful guideposts as RED undertakes its planning efforts for third-party evaluations of EHR programs.

As we have thought about what was learned from the "Footprints" effort and attempted to distill the main points from what was said in the papers, by the discussants, and by the general audience, we have identified two "messages."

- Message 1: There are a number of alternatives to the service delivery model that might be applied to EHR evaluations. Indeed, what we have referred to as the traditional model may be traditional in only a very limited context.

- Message 2: There are many different frameworks that can be used to evaluate EHR programs on which we have been focusing. The footprints we have started to uncover lead in many different directions. Before choosing a direction for any specific evaluation, the audiences for the evaluation and their general interests/concerns must be defined by EHR.

In the subsections that follow, we discuss these messages in somewhat greater detail. Specifically, we will examine the following topics:

- Who is the audience for EHR evaluations?

- Is there a set of core topics that all evaluations should address?
- What techniques are suitable for proposed evaluation tasks?

Who is the Audience for EHR Evaluations?

When the "Footprints" task was initiated, the audience for the evaluations was not identified and specific evaluation questions had not been spelled out. It is clear from the papers that participants had very different notions with respect to who the audience is or should be. For some, the audience was the personnel of projects that the programs had funded; for others, it was the educational research community; for still others, it was primarily Federal decisionmakers, including executive and congressional watchdogs and funding agencies. Some participants assumed that the evaluations had a narrowly defined accountability purpose, documenting the extent to which progress had been made toward the attainment of the short-term goals that projects had been set up to achieve. Others assumed that the evaluation should be guided by a heuristic perspective and assess the extent to which NSF programs had funded projects that dealt with important issues, had contributed to the generation of new knowledge, and could be expected to improve educational practice over time.

Several conference participants emphasized the need for audience definition before adopting the evaluation questions and methodologies that seem most appropriate. This point was strongly emphasized by two discussants with considerable experience in conducting federally sponsored evaluations (Raizen, Chelimsky), and was also addressed by several other participants (Johnson, David Jenness, Yin, Boruch).

Audience definition is also a question that RED, and not the research community, must ultimately answer. What are the questions that the upcoming cycle of evaluations are supposed to answer, and whose questions are they:

- The program directors', to tell them how well all or some of the program goals have been met?

- The NSF policymakers', to help them assess the relative effects of programs now in place and perhaps identify new directions for program priorities?

- The educational research community, to alert them to the results, dissemination, and footprints of work funded in the past and perhaps needed directions for future grant applications and grant reviews?

- Or administrators in NSF and in oversight agencies, to tell them which programs had the best effects (payoffs)?

Furthermore, the audience may or may not be the same for every evaluation that is to be undertaken. Before a final evaluation design is selected, the audience question needs to be answered since it is unlikely that a comprehensive evaluation, which would meet the needs and interests of all potential audiences, can be designed within current budget constraints.

Can a Standard EHR Evaluation Model be Developed?

In his overview of the "Footprints" conference, Scriven stressed the desirability of using a consistent model across the board for all programs funded by EHR, because this provides a degree of comparability. Stake, on the other hand, argued in favor of using different models depending on the structure and goals of each program. Webb also pointed to the need for using multiple methods of inquiry in light of the large number of variables and complexities characteristic in educational research. Furthermore, while the suggestions that emerged from the "Footprints" conference tended primarily, but not exclusively, toward qualitative approaches, several suggestions, particularly Yin's proposed analytic model, have a strong quantitative component. There are other ways in which quantitative approaches, such as sample surveys of project participants, e.g., teachers or administrators, could play a useful role.

The extent to which RED will decide to base its evaluation strategies for EHR programs chiefly on the suggestions of the "Footprints" conference participants depends of course on NSF's ultimate decisions about the target audience and judgments about the types of information that this audience will require. For example, if costs and benefits are to be an element that should be considered in the evaluation, evaluation models quite different from those proposed by the conference participants, incorporating quantitative approaches that were not mentioned would need to be developed.

While there can be no question that a standard evaluation model would have great advantages, we do not visualize how it can be implemented, given the diversity of programs and the likelihood that different audiences might be targeted for various types of program evaluations. However, we have concluded from the examination of common conference threads that there may well be a set of core evaluation topics and questions that can and should be included in all evaluations. These are discussed in the next subsection.

Ideas and techniques that RED should implement for all evaluations include:

- Tracking selected program footprints or impacts;

- Archiving utilization information;

- Using portfolio assessment;

- Exploring the role of intermediaries; and

- Examining timing and extent of dissemination.

Tracking Selected Program Footprints. Most participants found the "Footprints" concept a useful one, although for many of them, "Footprints" is primarily a tool to be used for the construction of more elaborate evaluation strategies. But as a first step in the implementation of any of the strategies recommended at the "Footprints" conference, a comprehensive and coherent inventory of existing footprints is needed.

Several of the presenters came up with long lists of evaluation questions that an examination of footprints could answer and suggested possible sources for locating them. (The paper by Boruch, who focused on the Studies and Indicator programs, was most specific with respect to the latter.) As suggested by the participants and discussants, these lists need to be reviewed, so that for each program, a manageable, preliminary list of footprints and their sources for each of the four "effects" areas outlined by RED (effects on the profession, on other research, on educational practices, and on the funding agenda of other institutions) can be established.

While such lists will no doubt be modified as the evaluation task progresses, it is imperative to start with the compilation of a systematic, well-defined, and parsimonious set of footprints for each program that is to be evaluated and documentary and other sources where these footprints might be located.

Several of the conference papers provide a good starting point for these compilations, but a good deal of additional work is required. Particular attention should be given to sources and informants that commonly used bibliographic searches will not uncover (see Boruch's suggestions). It is also likely that relevant information can be located in program and project files, for example in applications for grant renewals, progress reports, or peer reviews. Once a first set of footprints has been compiled, it may be productive to seek reactions and suggestions for additional types and sources of footprints from selected policymakers and researchers who are active in a given program area.

The next step must be the bounding, classification, and ordering of footprints, along conceptually meaningful dimensions. Thus, the accumulation and classification of footprint data is a complex task, requiring both the casting of a wide net to capture "hidden" footprints, the setting of boundaries, and the creation of "Footprints" categories that will enable the evaluator to perform meaningful descriptions and interpretations of the data. Whether or not boundary setting should precede the data collection, or be done subsequently, is probably best decided on a program-by-program basis.

Depending on the audience and design, this initial data compilation will provide the basis for the following evaluation activities:

- A crude assessment of the program's visibility and potential impact in each of the four "effects" areas mentioned earlier;

- The selection of outcome indicators and other variables for the construction of a causal model based on partial comparisons (Yin);

- The decision to substitute a sample of projects for the universe in order to carry out analytic procedures with a more manageable data set (Raizen's proposed methodology for sampling based on a project typology seems especially useful); and

- The selection criteria for case studies if the evaluation design calls for this activity.

Because the choice of evaluation strategies may be dependent to some extent on the volume and characteristics of footprints that are identified, NSF may find it useful to undertake the compilations prior to finalizing evaluation designs.

Archiving Utilization Information. As was stressed by Boruch and pointed out by several other participants, there is at this time no mechanism in place to obtain systematic information about the use of data and research findings generated by EHR. Knowledge resides at the program and project level in professional publications (citations, other references, etc.) and in public policy documents (minutes of congressional hearings, speeches by officials, etc.). To sustain an ongoing evaluation effort based on footprints, the establishment of an archive where this information can be stored and accessed is of great importance. In particular, program and project staff should be required to provide periodic "utilization information" to this archive.

Portfolio Assessment. Another recurring idea dealt with the need to take a broader perspective and look at the entire educational research system and at funding sources other than NSF when evaluating program effects. Also, rather than looking only at areas where footprints might be found, several authors and discussants identified a series of evaluation questions that would provide a meaningful context for footprints, suggesting some kind of mapping or portfolio approach:

- Is the universe of projects funded by EHR a true reflection of the interests of the research community (David Jenness)?

- What would have happened if projects other than those for which awards have been made would have been funded (Johnson)?

- Why are there no footprints from a funded project and what can be learned by looking at unsuccessful or unfunded research (Webb)?

While some of these questions, according to Johnson, call for the evaluator to measure the unmeasurable, it is evident that any evaluation of EHR programs would benefit from the more sophisticated approach of looking at EHR's "Footprints" programs in the broader context of the total science, mathematics, engineering, and technical education (SMET) research effort. This effort is funded by many sources besides NSF and carried out by researchers who have their own agendas, which influence how grant monies are expended and the extent to which performance bears a close relation to what was originally proposed in the funding applications (David Jenness, Boruch, Yin).

The questions raised by a number of participants addressed fundamental issues that the evaluation of the sizable and complex programs funded by EHR should consider:

- How well does each program target its awards?

- To what extent do programs address the right issues and respond to existing urgent needs for basic and applied research?

- Does the peer review process fund research stimulated by grantees' priorities for which they receive support from many sources?

- Do worthwhile proposals fail to obtain funding?

While NSF has instituted a mechanism for a broad review of these issues through periodic meetings of its Committee of Visitors and through the Expert Panels, a more systematic portfolio assessment is needed, based on an examination of funded awards, unfunded applications, funding activities carried out by other public and private agencies and an objective assessment of needs in the area for which the program bears responsibility.

One technique that might be useful in making portfolio assessments is a model proposed by Webb, represented on page 148, that uses a 2x2 matrix to address four key areas: what we have (or have not) learned from research supported by a program, the extent to which findings have been used, what problems have not been addressed by the program, and how the gap was filled. Webb limited himself to the RTL program when he developed this model and proposed specific types of studies for answering the questions raised. However, the model could be adapted for all or most EHR programs, since it goes to the core of issues that concern educational leaders as well as policymakers in funding and oversight agencies.

The Role of Intermediaries and Gatekeepers. Several of the papers have pointed to the important role played by intermediaries in acting as facilitators and gatekeepers in acquainting potential users (policymakers and practitioners) with research findings. Although this issue relates to some extent to dissemination, it should be examined in the "Footprints" context and needs to be considered for every EHR program that is being evaluated, although the types of intermediaries and the gatekeeping function they perform will differ widely.

In her paper, Christine Dwyer argued for a full-blown study of the paths and processes by which the Research in Teaching and Learning Program (RTL) influences educational practice, by examining the treatment of NSF-generated information by intermediaries and exploring the factors that determine transfer/nontransfer of this knowledge to practitioners (school personnel). The case studies that Dwyer proposes as a first step are exploratory in nature, focusing primarily on the intermediaries modus operandi, rather than systematic attention to the fate of EHR products. In her discussion, Raizen raised several caveats. In particular, she cautioned that intermediaries must be carefully selected and that not all intermediaries afford a valid test of information exchange. She also felt that rather than using the policies and practices of intermediaries as the starting point for case studies, it might be more useful to start out with some specific practice that looks as if it had been influenced by some assessed program and then trace back where the practice came from. Another approach that NSF may want to consider is to look at one major project within a given program to examine its treatment by relevant intermediaries (including some, such as museums,

Exhibit 1

Research Results	Applications	
	Yes	**No**
Know	What findings and information have been produced that have successfully solved a problem or fulfilled a need?	What findings and information have been produced that have not been applied to solve an important problem or fulfill a need?
Do Not Know	What critical problems or needs have not been resolved or refined by research findings and information?	What negative or poor applications have filled the gap in the absence of solid research findings and information?

whose main function is not service to education practioners), and examine the extent to which its findings did or did not reach the targeted audience. If carefully shaped so as to focus attention on the issue of concern to EHR, pilot studies of the role played by intermediaries could be very useful indeed.

Dissemination. There can be little argument that in many cases, the number of footprints is directly related to dissemination efforts on the part of investigators. NSF may want to investigate the extent to which the footprints that have been uncovered resulted from dissemination efforts by NSF program and project staff, and identify those dissemination techniques that have been most effective in yielding footprints. Initially, one or two case studies might be undertaken.

The many related issues, which the conference participants touched upon but did not develop, addressed the relationship between evaluation and dissemination. Several discussants (Raizen, Chelimsky, and Scriven) pointed out that dissemination is not appropriate for all research undertakings and is an expensive activity. Hezel, on the other hand, felt that evaluating the dissemination activities was a major task for the evaluation. There was also no thorough discussion about how to reconcile the need for early and widespread dissemination, which is emphasized in NSF proposal guidelines, with the time constraints imposed by evaluation and validation of project results, when projects are designed to affect educational practice and replication of successful projects is a program goal.

Scriven stated in his summation that although dissemination was included in the presentation and discussion of several conference participants, it was not a topic on the "Footprints" agenda and should be treated as an important but separate topic from evaluation.

Ideas and techniques that may differ with respect to various evaluations include:

- Need for causal attribution;
- Choice of evaluation methodology;
- Use of innovative analytic frameworks; and
- Use of innovative data collection.

Need for Causal Attribution. Those participants who tended to focus on the evaluation needs of Federal stakeholders (NSF, OMB, and Congress) and on the harder question of program worth felt that causal attribution had to be an essential ingredient of evaluations of federally funded programs (Scriven, Raizen, Chelimsky). In some cases impact attribution may also be important for program and directorate staff or the educational research community; in other cases, it may be more useful to devote resources to more extensive descriptive data for these audiences. The question of causal attribution was most fully addressed by Yin, who devoted his paper to the presentation of a new analytic technique to assess program effectiveness and make possible causal attribution of effects in the absence of controlled evaluation designs. Webb's paper also addresses the issue of attribution of effects. The recommendations of Yin and Webb are discussed in greater detail below (analytic frameworks).

Choice of Evaluation Methodology. In setting out the "Footprints" task, RED emphasized the need for finding new ways of evaluating the unique and innovative programs being supported in mathematics and science education and suggested that both new methodologies and new questions needed to be developed. While the participants presented many different ideas and differed on many issues, the one point on which there was agreement among the largest number of presenters and discussants was that the prevailing educational evaluation methodology, the service delivery model, is inadequate for the evaluation of many EHR programs and that viable alternatives do exist.

The alternatives offered took on many dimensions. At times nontraditional was equated with qualitative, and, therefore, traditional was associated with quantitative methods. Some participants (Barley and Mark Jenness) defined nontraditional methods as those that emphasize the interests of project clients and other local stakeholders and use negotiation as the major evaluation tool. While Stake questioned the use of any systematic evaluation method (because of the dominance of the political and administrative context in which the programs operate), most participants offered nontraditional evaluation strategies using both improved new approaches to educational evaluation and traditional scientific methods from other fields, especially ethnographic and cultural studies.

Indeed, the description of proposed nontraditional approaches led one discussant (Phelps) to comment that "they all model what should be and is good evaluation practice. They are only nontraditional in the sense that in the Federal Government they are not often carried out."

In his comments, Scriven took exception to the widely expressed need for new methodologies. In his words, he found himself in the unfamiliar position of defending the status quo. He felt that the arguments for needing a new approach were mostly aimed at what is really a straw man and faulted the NSF's procurement policies, rather than shortcomings of the methodology. He asserted that the agency had not tapped into the best available evaluation practices, which are a kind of eclectic amalgam of qualitative and quantitative methods, carried out by experienced and sophisticated evaluators.

Taken together, the comments by conference participants suggest that while RED should continue to encourage the development of innovative methodologies, there is no need to rely solely on methodologies developed from scratch. While it may be necessary to do so for the evaluation of some programs, for others (for example the RTL program) the "eclectic mix" recommended by Scriven may be most appropriate. Furthermore, there presently exists a number of fully or partially developed models that are not based on the service delivery approach. A first step should be to explore the alternatives with the goal of adopting (or adapting) some of the quantitative and qualitative approaches that already are used in our own and other fields. The ideas and techniques proposed by the "Footprints" authors may be considered nontraditional with regard to common practice in federally funded evaluations, but many of them are based on data collection and analytic approaches with established histories and credibility.

Alternative Analytic Frameworks. Three of the conference papers (Yin, Webb, Barley and Mark Jenness) focused on innovative techniques for developing analytic frameworks for EHR evaluations.

Yin's objective was to use footprints to establish a causal link between program activities and observed outcomes through the use of a rigorous technique that would be an acceptable substitute for experiments or quasi-experiments used in traditional service delivery-based models, which are inappropriate for most EHR programs. The usual characteristics of grant programs are that the intervention carried out by grant-funded projects is weak or small, relative to the impact of interest; the intervention is not part of a formal research design; and extensive time or resources are not available for the research effort. Given these problems, experimental designs must be ruled out. Database analyses are primarily descriptive and do not permit causal inferences.

Instead, Yin recommends a new methodological strategy, which aims at making "multiple, partial comparisons" instead of imposing a singular research design in carrying out an evaluation. Unlike traditional evaluation designs, this method can be used when evaluators have no control over the intervention or when the interventions do not meet the statistical requirements of any of the "traditional" designs. Partial comparisons can enable investigators to offer causal inferences by using single components (specific project effects) as the main unit of analysis. The larger the number of positive inferences that can be supported through these partial comparisons, the stronger the argument that positive results were produced and the stronger the conclusion that the program under evaluation produced them. This strategy requires the evaluator to identify and collect data, in effect footprints, that can satisfy as many partial comparisons as possible. Outcome data from projects funded by the program are the relevant input for each partial comparison, and the instruments needed to collect these data will vary. The AAT program was one for which he felt this approach would be especially suitable.

The paper presented by Webb presented several strategies for the analysis of footprints. Especially useful was his suggestion about dealing with the very large number of footprints that some programs are likely to yield (he focused on the RTL program that to date has funded more than 200 projects). One of the issues often raised by critics of qualitative approaches is that investigators are very good at collecting a great deal of interesting data but have not developed rigorous methodologies for their interpretation. Webb proposed a generalizability analysis to substitute the study of a sample of projects, selected at random, that would yield a cross-section of projects and provide a good description of the program as a whole. In her discussion, Raizen proposed an alternative to random sampling of projects, recommending instead a two-stage approach, with some initial grouping of projects along common dimensions, such as problem addressed, or approach taken, and subsequent sampling within each of these groups. Raizen emphasized that the groupings would have to be thought through very carefully, but if this was done, the sample used for analysis would be greatly superior to one obtained through random sampling.

Both Webb and Yin sought to build comprehensive evaluation models to shed light on the value of programs, address the issue of utilization of findings, and answer questions of causality. Webb's approach, discussed earlier, used a 2x2 matrix to examine the extent to which research has yielded findings that were used to solve educational problems. Yin's model incorporated the concept of rival hypotheses to test the causal link between research findings and the adoption of educational innovation. His proposed analytic technique, partial comparisons, appears promising. Considerable work on partial comparisons has already been done by Yin for other agencies.

The framework proposed by Barley and Mark Jenness is based on a different premise. They believe that the main goal of evaluation is formative and aimed at project and program improvement. Their proposed cluster evaluation concept and techniques for its implementation have been tested, with support from the W.K. Kellogg Foundation for formative but not for summative assessments. Barley and Mark Jenness recommend its use for summative program evaluation through the creation of samples of retrospective clusters, consisting of completed projects, based on regional or topical sampling frames. A "cluster evaluator" would work with directors and other project staff to negotiate a set of common cluster outcomes and collect both qualitative and quantitative data from a variety of sources using various techniques. Some common cluster instruments, used across projects to collect consistent data, can be created for the data collection. Scriven has forcefully argued against this approach, pointing to the credibility and objectivity issues that its use would create for a summative evaluation. A more limited use of this technique, confined to data collection only and discussed later in this paper, might be considered.

Incorporation of all or part of Webb's and Yin's models and techniques in an evaluation design would greatly increase the sophistication of footprints analyses. Both models would require substantial data collection, in particular a fairly complete mapping of all efforts sponsored by public and private agencies that are directed at the strengthening of mathematics and science education and recruitment. This mapping would be a difficult and time-consuming undertaking; again, a sampling approach seems indicated. After data have been collected, the suggested models for attributing specific outcomes to EHR programs can be fleshed out.

Yin sees the need for further methodological development before the partial comparisons technique can be tried for the evaluation of NSF programs. Key outcome measures (for example, new ideas for research or practice) have to be developed. To pinpoint effects traceable to NSF-funded programs, case studies need to be conducted of funded investigators and the projects they undertake, so as to develop information about how grantees merge various sources of support to carry out their research projects. The list of partial comparisons needs to be expanded to be suitable for EHR programs, and pilot testing should be done to assess the efforts and costs required. But if EHR sees the need for in-depth assessments of program outcomes, these methods are certainly worth exploring further.

Innovative data collection. Several of the papers, especially those by Boruch, Johnson, and Barley and Mark Jenness, contain innovative suggestions for data sources and data collection techniques that could be explored. Boruch, who focused his discussion on RED's Studies and Indicators programs, offered an extensive list of possible sources of references and uses going beyond the commonly used citation counts and publications in refereed journals by high-quality publishers. He suggests professional recognition through awards and prizes, presentations in professional and public forums, and popular press or media coverage. He also recommends scanning press and agency reports that have used a study without directly acknowledging the source, direct observation of public meetings where studies are discussed, and self-reports by project staff, usually the principal investigator. Peer reviews, review panels, and the knowledge of seasoned staff in foundation grant programs and Federal agencies are other good sources. Boruch further pointed to somewhat more remote effectiveness indicators, such as contributions to research methodology and data production methods. He recognizes that the systematic accumulation of this information may well be a monumental task, best carried out in an academic setting where graduate students constitute an affordable labor source.

Clusters could be a practical data and information collection resource, standardize evaluation questions. The RTL program is a good candidate for this approach. Using a common data collection instrument for projects in a given cluster would standardize evaluation questions and facilitate the collection of a common core of data for a given program. This approach might be useful for the RTL program.

Recommendations

The reason for initiating the "Footprints" task was to develop some nontraditional approaches to evaluating EHR programs, which, because of their organization, goals and support structure, are not easily amenable to being examined using the typical Federal evaluation model. The varied experts whose ideas were tapped as authors or discussants have provided NSF with a long list of ideas from which to choose in approaching these evaluations. In this paper, we have selected for more extensive discussion those suggestions that we felt were especially promising. While several useful methodologies and frameworks for assessing programs' worth have been offered, we believe that the most useful contribution that the conference (and this paper) may have made is the identification of the common core of activities that we have outlined: tracking selected program footprints, portfolio assessment, the role of intermediaries, and the relationship between evaluation and dissemination. We also feel that the identification of evaluation audiences is of paramount importance before specific evaluations are designed.

What happens next depends on a number of steps that EHR itself must take; steps that involve possibly investing in the fuller development of some of the alternatives offered, as well as setting priorities among audiences and questions to be addressed. Given the innovative nature of some of the proposed procedures, small-scale pilot testing would also be advisable. We have identified several techniques that EHR may want to consider in planning upcoming evaluations for specific programs, and some methodological tasks that might be undertaken prior to the adoption of final evaluation designs. These include:

Develop a System for the Collection of Footprints from NSF Program and Project Files. Several discussants pointed to the role that NSF itself, as well as funded projects, must play in accumulating footprints. These recommendations have been discussed earlier. Written requests for copies of reports and other types of information, telephone inquiries about findings, invitations extended to program and project staff to participate in activities where program-generated information is to be discussed are not systematically documented at the program level. Boruch saw the need for an NSF program archive; other presenters emphasized the role of the project director. At present, available information is largely anecdotal and decentralized. As part of the current EHR effort for database creation, it may be possible to generate systematic Footprint data at the program and project level.

Develop a Methodology for Portfolio Assessment. The Webb matrix represents one possible approach; Yin's "rival hypothesis" also addresses the issue. But EHR needs a comprehensive strategy to carry out this assessment for all its programs.

Conceptualize and Pilot-test the Intermediary Function as it May Apply to all EHR Programs. Once appropriate intermediaries have been identified for several programs, it may be useful to adopt Raizen's strategy and examine in a pilot test the role played by these intermediaries with respect to one or more products that resulted from these programs.

Clarify EHR's Policy with Respect to the Connection between Evaluation and Dissemination. Here, too, it would probably be useful to look at some actual dissemination practices and examine their effectiveness as well as their relation to evaluation efforts and outcomes.

If the Causal Attribution of Program Effects is to be Included in the Evaluation, Develop and Pilot-test the Partial Comparison Methodology for the Program to be Evaluated. As suggested in the earlier discussion, it is not obvious that the model and analyses proposed by Yin will be appropriate for all EHR evaluations. When they are used, considerable methodological development and pilot testing will be needed, as Yin himself has emphasized.

Index

A

- Aaron, H., 116
- Accountability, 107, 111
- AERA, 22
- American Evaluation Assoc., 107, 133
- Application of Advanced Technology (AAT) Program
  - description, 3, 45;
  - evaluation criteria, 32-35;
  - funding level, 26;
  - goals, 27, 47;
  - "high risk" portfolio, 28, 45;
  - nature of, 140;
  - program pay-offs, 32, 137;
  - proof of concept, 32
- Assessment
  - of learning, 18;
  - of needs, 43, 91;
  - of program progress, 28
- Auletta, K., 121

B

- Ball, D.L., 111
- Barley, Z.A., 103, 105
- Barley, Z.A., & Jenness, M., 77, 127, 129
- Barnouw, E., 110, 112
- Bell, T., 113
- Bellavita, C., Wholey, J.S., & Abramson, M.A., 99
- Booz, Allen, & Hamilton, 121
- Borg, W.R., & Gall, M.D., 59
- Boring, E.G., 118
- Boruch, R., 41
- Bruner, 41
- Bureau of Labor Statistics, 13

C

- Carpenter, T.P., & Moser, T.M., 60
- Carroll, J.B., 61
- Chelimsky, E., 110, 120, 131, 133
- Cluster analysis. *See* evaluation methodology
- Coles, R., 111
- Collaboration
  - across agencies, 12;
  - across projects, 21
- Cook, T.D., 120
- Cronbach, L.J., 99, 105, 107, 108, 110, 115, 119, 120, 122

D

- Data. *See also* evaluation methodology; impact measures
  - acknowledgment of source, 9;
  - bounding collection of, 79, 88, 93, 145;
  - case histories as collection strategy, 85;
  - in cluster analysis, 103;
  - collection of, 46, 89, 152;
  - enhancing use of, 12;
  - by museums, 91;
  - qualitative, 43, 50, 93-94, 119;
  - quantitative, 1, 26, 43, 50, 93-94;
  - sharing by government agencies, 12;
  - tracking use of, 7-8
- Denzin, N.K., 117
- Denzin, N.K., & Lincoln, Y., 112, 116
- Department of Education
  - Eisenhower program, 20, 37, 76, 77;
  - evaluation activities, 37;
  - National Center for Education Statistics (NCES), 7, 10, 13;
  - National Diffusion Network, 76, 77, 79, 94;
  - OERI, 76;
  - Planning and Evaluation Service, 7
- Dissemination
  - of AAT project outcomes, 45, 46;
  - audience for, 94;
  - in cluster evaluation, 105;
  - and data use, 12;
  - evaluation of, 48, 75-76;
  - and footprints, 82, 148, 153;
  - need for, 45, 93, 97, 127-128, 131-132;

(Dissemmination *continued*)
- role of intermediaries, 79

Donmoyer, R., 99

E

Eisenhower program. *See* Department of Education

Eisner, E., 115

Erickson, F., 116

Evaluation. *See also* assessment; evaluation methodology
- and accountability, 54, 107, 111;
- and advocacy, 98, 107;
- audiences, 49, 53, 93, 127, 143-144;
- cost/budget considerations, 54, 109, 144;
- of dissemination, 48;
- and educational practice, 19, 54, 98;
- intergovernmental cooperation, 37-38;
- need for explanation and understanding, 114, 132;
- and policy, 12, 18, 33, 108;
- and politics, 110-111, 128;
- program and project, 2, 27, 43, 97, 140;
- purpose of, 97-98, 110, 132;
- questions, 19-21, 31, 39, 48, 50-52, 62-66, 77-78, 100-101, 115-116, 127;
- relation to activities not funded by NSF, 32, 35, 40, 41, 146-147;
- requirements for small grants, 129;
- theory, 132

Evaluation methodology. *See also* data; evaluation; impact measures; validation
- application matrix, 62-63;
- audit model, 121;
- causal inference, 1, 29, 31, 108, 110, 134, 148, 152;
- expert panel, 28, 137-138;
- formal vs informal, 108;
- generalizability analysis, 55, 71-72, 73, 95, 151;
- goal based vs goal free, 47, 48, 52;
- meta-analysis, 136-137;
- non-traditional, 1, 2, 22-23, 112, 114, 149-150;
  - based on interpretation, 115-116, 119-122;
  - cluster analysis, 77, 89, 97, 101-106, 127, 129, 135, 151-152;

(Evaluation methodology, nontraditional *continued*)
  - constructivist, 48, 114, 117;
  - empowerment, 135;
  - naturalistic, 47, 50, 114;
  - subjectivist, 47;
- partial comparisons, 29-32, 146, 150-151, 153;
- and program goals, 141;
- product vs process, 110, 114;
- qualitative vs quantitative, 50, 112, 119, 127, 128, 133, 149-150;
- research community culture analysis, 54, 70-71, 73, 95;
- retrograde analysis, 54, 66-67, 95;
- retrospective vs prospective, 46;
- sampling, 72, 95, 146;
- standard model, 135, 144;
- statistically significant differences, 136;
- traditional, 1, 15, 25-26, 98, 115, 133, 140;
- triangulation, 117;
- use of multiple methods, 61, 108, 128;
- use of social problem study group, 99, 120;
- video documentary, 54, 67-70, 73, 95

Evaluation Network, 134

F

Fennema & Carpenter, 57

Fitzsimmons, S.J., 35

Footprints
- absence of, 16, 20, 146;
- archive, 146;
- critique of concept, 108-109;
- definition, 1-2, 7, 142;
- and dissemination, 48, 82;
- and educational reform, 78;
- as evaluation tool, 79, 142;
- of Indicator and Studies programs, 7;
- origin of metaphor, 3;
- rationale, 15;
- of RTL program, 19, 21, 22;
- sources of, 8-11, 50, 82, 142, 145, 153

G

Geertz, C., 116

General Accounting Office (GAO), 8-9, 13, 29, 120

Gordon, S.W., and Bhattachanyya, 17

Grant process
  and documentation of research use, 10;
  and evaluation methods, 25, 27-28, 141;
  grantees' agendas, 25, 28, 41;
  need for documentation of activities, 41;
  philosophy, 27-28, 141
Guba, E.B., 77, 88
Guba, E.B., & Lincoln, Y., 99, 100, 108, 114, 115
Guiton, S., & Burstein, L., 109

H

Hattie, J., 136, 138
House, E.R., 49

I

Impact measures
  and advocacy, 107;
  conceptualization, 32;
  direct vs indirect, 109;
  in evaluation models, 26, 33-34, 127, 135;
  indicators, 2, 49;
  methodological innovations, 11;
  new ideas as, 32;
  obstacles in NSF programs, 26, 28;
  and political pressure, 107;
  qualitative vs quantitative, 52, 93;
  in relation to goals, 47;
  sources of information, 48;
  unanticipated outcomes, 47, 48;
  validity of, 109, 111-112
Implementation 75, 80
Indicators Program
  assessing data use, 7-13;
  description, 3, 7;
  footprints, 7-11;
  and policy making process, 13;
  program goals, 32
Informal science education, 91
  and use of NSF research products, 92
Intermediary organizations, 41
  description, 76;
  and implementation, 75;
  and knowledge transfer, 9, 77, 78, 79;
  museums as, 91;
  and program evaluation, 77, 94, 153;
  role in education reform, 76, 147-148;
  selection of, 88, 94
International Mathematics Study
  second, 11, 58, 68;
  third, 7
Interpretation council, 120-121, 127, 137;
  and External Expert Panel, 120

K

Kaplan, A., 60
Katzenmeyer, C., 119, 120
Kellogg Foundation, 101
Kelly, M., 111
Knowledge utilization, 12, 80
Krueger, M., 112

L

Levin, 136
Louis, K.S., 79
Lincoln, Y. & Guba, E.S., 105
Lindblom, C.E. & Cohen, D.K., 122

M

McGrawth, J.L., 32
McLaughlin, M., 80
McLuhan, M., 67, 113
Miles, M.B. & Huberman, M.A., 119
Mosteller, F., 136, 138

N

National Center for Education Statistics (NCES). *See* Department of Education
National Council of Teachers of Mathematics (NCTM), 22, 57, 60, 86, 111
National Diffusion Network (NDN). *See* Department of Education
National Research and Education Network (NREN), 9
National Science Foundation (NSF), 8, 9, 13, 15, 21, 27, 28, 40, 43, 77, 78, 133
Nelson, C.E., 80

O

Office of Inspector General (OIG), 133, 137
Outcome measures. *See* impact measures

P

Peer review, 1, 122
  and project funding, 95, 137;
  and use of information, 10

Piaget, J., 60
Pilot testing of innovative methodologies, 153
Polanyi, M., 113
Porter, A.C., 110
Pskotka, J., 113
Principal Investigator (PI)
  as source of information, 10, 48;
  career as outcome indicator, 95
Project evaluation. *See* evaluation

R

Reform in mathematics and science education, 18, 59, 78, 81, 83
Research in education
  basic vs applied, 59;
  and educational practice, 61;
  importance of theories and models, 60-61;
  and museums, 91;
  relation between research and evaluation, 132;
  science and math, 61, 130;
  typology, 59;
  "uses of," 7
Research in Teaching and Learning Panel, 15, 21
Research in Teaching and Learning Program (RTL)
  and cluster analysis, 98;
  evaluation questions, 19-21, 62-66, 77-78, 94, 132, 134;
  footprints, 19, 21-22;
  goals, 18-19, 40, 55-56;
  nature of, 95, 140;
  program description, 3, 17-18, 55, 57-58;
  use of generalizability analysis, 95
RMC Research Corp., 81
Rockefeller Foundation, 7, 12
Rossi, P., 41

S

Schwandt, T., 114, 117
Shavelson, R.J., 109
Scriven M., 47, 107, 109, 110, 115, 122
Smith, N.L., 77, 88, 134
Spiro, R.J., 112
Stake, R.E., 115, 118, 127, 129, 130, 132, 133, 136, 137
Stake, R.E. and Trumbull, D., 113, 116
Stakeholder, 49-50;
  as clients, 49;
  and program effects, 48;
  role in evaluation, 2, 82, 97, 99;
  as sources of information, 47
Stevenson, H., 58, 72
Strauss, A. and Corbin, J., 120
Studies Program
  assessing data use, 7- 13;
  definition, 3, 7;
  footprints, 7-11;
  funding level, 26;
  goals of, 26, 32;
  nature of, 140;
  and policy making process, 13

T

Technical Assistance Centers (TAC) 76, 77, 85-88
Technical assistance agencies, 79
Trumball D., 114

U

Underrepresented groups
  empowerment for, 56;
  program goals and impacts for, 18, 20;
  recruitment for professional activities, 20

V

Validation. *See also* evaluation methodology
  contradictory results, 54;
  of effectiveness indicators, 111, 113;
  of observations, 117;
  of program goals, 47;
  of traditional evaluation methods, 128
Van de Vall, M., 80
Von Wright, G.H., 114
Von Maanan, J., 117

W

Webb, N.L., 89, 91, 95
Webb, N.L. Schoen, H. and Whithurst, S.D., 67
Weick, K., 89
Weiss, C., 79
Wholey, J.S., 98
Wholey, J.S. and White, B.F., 98
Wittgenstein, 132

Worthen, B.L. and Sanders, J.R., 46
Woolley, B., 112

Y

Yin, R.K., 29, 89, 119
Yin, R.K. and Sivilli, T.S., 29

Z

Zacharias, 41

NATIONAL SCIENCE FOUNDATION
ARLINGTON, VA 22230

OFFICIAL BUSINESS
PENALTY FOR PRIVATE USE $300

RETURN THIS COVER SHEET TO ROOM P35 IF YOU DO NOT WISH TO RECEIVE THIS MATERIAL ☐ , OR IF CHANGE OF ADDRESS IS NEEDED ☐ , INDICATE CHANGE INCLUDING ZIP CODE ON THE LABEL (DO NOT REMOVE LABEL).

SPECIAL FOURTH-CLASS RATE
POSTAGE & FEES PAID
National Science Foundation
Permit No. G-69

3

00145173 ERIC
ERIC FACILITY
1301 PICCARD DRIVE
SUITE 300
ROCKVILLE MD 20850-4305

BEST COPY AVAILABLE

NSF 95-41 (new)